Mente come Inferenza: Il Principio di Energia Libera tra Neuroscienze Computazionali e Filosofia Analitica

Il Problema e la Promessa

I. Preambolo: un nodo che non si scioglie

Esiste nella storia della filosofia un problema che si rifiuta ostinatamente di cedere alle soluzioni, che resiste all'erosione del progresso scientifico con la stessa tenacia con cui una roccia resiste all'onda, assorbendola e restando immobile. Il problema è quello del rapporto tra mente e cervello — o più precisamente, nella formulazione che David Chalmers ha reso celebre nel 1995, quello di spiegare perché ai processi fisici e computazionali del sistema nervoso si accompagni qualcosa che è come essere quel sistema: un'esperienza soggettiva, un vissuto qualitativo che sentiamo irriducibile alla sola descrizione funzionale.

Chalmers distingueva tra i problemi «facili» e il problema «difficile» della coscienza. I primi riguardano funzioni cognitive che, per quanto complesse, sembrano in linea di principio spiegabili nei termini di meccanismi computazionali: l'integrazione dell'informazione, il controllo del comportamento, l'accesso introspettivo agli stati interni, la discriminazione percettiva. Il secondo — il problema difficile — chiede invece perché questi meccanismi siano accompagnati da qualia, da quel carattere fenomenico per cui vedere il rosso di un tramonto è qualcosa di più e di diverso dall'elaborare informazione sulla lunghezza d'onda di circa 700 nanometri. Questo divario esplicativo, questo explanatory gap che Joseph Levine aveva già intuito nel 1983, sembra impermeabile alle spiegazioni in terza persona.

Negli ultimi due decenni, tuttavia, una proposta teorica di straordinaria ampiezza e ambizione ha guadagnato terreno tanto nelle neuroscienze quanto nella filosofia della mente: il Principio di Energia Libera (Free Energy Principle, d'ora in avanti FEP), elaborato con rigore matematico crescente dal neuroscienziato Karl Friston a partire dai primi anni 2000. Questa proposta non è soltanto una teoria della percezione, né soltanto un'architettura computazionale del cervello: nella sua formulazione più radicale, aspira a essere un principio primo dell'esistenza stessa dei sistemi adattativi dotati di un confine identificabile, un'ontologia della mente che si articola attraverso la nozione di inferenza attiva.

La domanda che questo saggio si propone di affrontare è dunque duplice, e le due parti si implicano a vicenda. In primo luogo: cosa afferma esattamente il FEP, nella sua articolazione

matematica rigorosa, e perché costituisce un avanzamento genuino rispetto alle precedenti teorie computazionali del cervello? In secondo luogo, e qui risiede la tesi originale che si intende difendere: il FEP riesce là dove altri approcci hanno fallito, ossia nel dissolvere o risolvere il problema difficile della coscienza? La risposta che si argomenterà è negativa, ma in modo non banale: il FEP non dissolve il problema, bensì lo trasforma profondamente, ricollocandolo all'interno di una struttura matematica che chiarisce con precisione inedita dove e perché l'abisso esplicativo persiste — e, facendo ciò, indica le condizioni che una futura teoria della coscienza fenomenica dovrebbe soddisfare.

II. La tradizione che precede: dal comportamentismo al bayesianesimo

Per apprezzare la portata del FEP occorre situarlo nel paesaggio intellettuale da cui emerge. Il tentativo di naturalizzare la mente — di spiegarla integralmente nei termini della scienza naturale — ha attraversato diverse fasi che vale la pena di richiamare non in forma enciclopedica, ma in modo da isolare i nodi concettuali che il FEP prende di petto.

Il comportamentismo, nella sua versione filosoficamente più sofisticata, tentava di eliminare il vocabolario mentale traducendolo integralmente in termini di disposizioni al comportamento. Gilbert Ryle, nel 1949, sosteneva che parlare di stati mentali non è parlare di eventi privati che avvengono in un «fantasma nella macchina», bensì descrivere le tendenze comportamentali di un organismo. Questo approccio aveva il merito della parsimonia ontologica, ma pagava un prezzo altissimo: lasciava inesplicata la caratteristica più saliente dell'esperienza soggettiva, la sua qualità intrinseca che non si riduce a nessuna disposizione disposizionale. Come diceva Thomas Nagel nel suo celebre saggio del 1974, anche se conoscessimo ogni disposizione comportamentale di un pipistrello, non sapremmo ancora cosa prova quel pipistrello quando si orienta con l'ecolocalizzazione.

La rivoluzione cognitivista degli anni '50 e '60 sembrò aprire una via d'uscita introducendo il livello computazionale come livello esplicativo autonomo. Nella tripartizione di David Marr — livello computazionale, livello algoritmico, livello implementazionale — il problema della mente veniva ricondotto alla specificazione di cosa calcola un sistema, come lo calcola, e su quale substrato fisico. Questa architettura concettuale era elegante e produttiva, ma lasciava irrisolto il problema di come il livello computazionale potesse dar conto del livello fenomenico: la distinzione di Marr riguarda l'elaborazione dell'informazione, non l'esperienza di quell'elaborazione.

Il funzionalismo, nella sua versione più influente elaborata da Hilary Putnam, sosteneva che gli stati mentali sono stati funzionali: sono individuati dai loro ruoli causali rispetto agli input sensoriali, agli output comportamentali e agli altri stati mentali. La realizzabilità multipla — il fatto che lo stesso stato funzionale possa essere implementato su substrati fisici diversi — sembrava assicurare l'autonomia della psicologia rispetto alla neuroscienza. Anche qui, però, il problema difficile sfuggiva: i critici, a cominciare da Ned Block con la sua distinzione tra access consciousness e phenomenal consciousness, mostravano che la coscienza di accesso (la disponibilità di un'informazione per il ragionamento e il controllo del comportamento) non implica logicamente la coscienza fenomenica.

Il bayesianesimo cerebrale — la proposta che il cervello implementa l'inferenza bayesiana per costruire rappresentazioni probabilistiche del mondo — ha rappresentato un passo fondamentale verso la formulazione del FEP. Nelle sue radici storiche si riconosce Hermann von Helmholtz, che nel 1867 descriveva la percezione come «inferenza inconscia» (unbewusste Schlüsse), un processo per cui il sistema nervoso genera ipotesi sul mondo esterno e le verifica contro i dati sensoriali. Nel Novecento, Richard Gregory aveva sviluppato questa intuizione in una teoria della percezione come costruzione ipotetica, e negli anni '80 e '90 la statistica bayesiana aveva fornito gli strumenti formali per rendere questa intuizione matematicamente precisa.

La teoria del codice predittivo (predictive coding), sviluppata da Rao e Ballard nel 1999 e da altri successivamente, articolava un'architettura gerarchica in cui ogni livello della gerarchia corticale genera predizioni sui livelli inferiori e trasmette verso l'alto soltanto gli errori di predizione — le discrepanze tra predizione e input effettivo. Questa architettura non era soltanto computazionalmente elegante: trovava riscontro in dati anatomici e neurofisiologici riguardanti le connessioni feedback e feedforward nella corteccia visiva. Ma rimaneva una teoria della percezione, ancorata a specifiche architetture computazionali e non ancora elevata al rango di principio unificante.

III. La promessa del Principio di Energia Libera

Il contributo di Friston, a partire da una serie di articoli pubblicati tra il 2005 e il 2010 e poi sviluppati in lavori sempre più ambiziosi, consiste nell'unificazione di questi fili disparati sotto un unico principio variazionale. La tesi di fondo è che qualsiasi sistema che mantenga la propria organizzazione nel tempo — che resista alla tendenza all'entropia propria dei sistemi fisici aperti — deve necessariamente minimizzare una quantità chiamata energia libera variazionale. Questa minimizzazione, applicata agli stati interni del sistema e alle sue azioni sul mondo, è equivalente a compiere inferenza bayesiana approssimata sul proprio ambiente.

Detto così, il principio suona come una generica parabola sull'adattamento. Ma la sua forza sta nella specificità matematica con cui il concetto di energia libera è definito, nelle sue molteplici decomposizioni equivalenti, e nelle conseguenze che ne derivano per la struttura dell'inferenza percettiva, dell'azione e, in ultima analisi, dell'identità stessa di un sistema come sistema — come entità distinta dal proprio ambiente.

Tre sono le promesse che il FEP avanza, e che questo saggio esaminerà criticamente.

La prima è una promessa di unificazione. Il FEP vorrebbe riconciliare, sotto un unico formalismo, la percezione, l'azione, l'apprendimento e l'attenzione come aspetti diversi di un'unica strategia di minimizzazione dell'energia libera. Questa unificazione non è meramente descrittiva: è normativa nel senso che specifica cosa i sistemi adattativi devono fare per esistere come tali, e permette di derivare da questa norma le architetture computazionali specifiche — come il predictive coding — come casi particolari.

La seconda è una promessa di naturalizzazione della mente come processo inferenziale. Se la percezione è inferenza, se l'azione è inferenza attiva, se persino il sé è una specie di modello

inferenziale, allora si delinea la possibilità di una teoria della mente interamente naturalizzata, radicata nella fisica statistica e nella teoria dell'informazione, senza residui mentalisti.

La terza — e più controversa — è una promessa di dissolvere o aggirare il problema difficile della coscienza. Alcuni sostenitori del FEP, tra cui Friston stesso in certi passaggi, suggeriscono che la distinzione tra ciò che è come essere un sistema e il sistema inteso come processo fisico sia in qualche modo tracciata dalla struttura stessa del confine di Markov, la frontiera statistica che separa un sistema adattativo dal suo ambiente. È su questo punto che si concentrerà la critica più penetrante.

IV. La tesi: trasformazione senza dissoluzione

La tesi che si argomenterà in questo saggio può essere formulata nel modo seguente. Il FEP costituisce un progresso genuino e fondamentale nella comprensione della mente come processo computazionale e inferenziale. Esso fornisce un'architettura matematicamente rigorosa entro cui la percezione, l'azione e l'apprendimento trovano un'unificazione elegante, e offre predizioni verificabili sulla struttura del cervello che trovano crescente supporto empirico. Tuttavia, il FEP non dissolve il problema difficile della coscienza, né può farlo in linea di principio: esso opera interamente al livello della coscienza di accesso, della struttura funzionale e computazionale dei processi mentali, e non tocca la questione di perché a questa struttura sia associata un'esperienza soggettiva qualitativa.

Ciò che il FEP fa, e non è poco, è precisare le condizioni formali che un sistema deve soddisfare per essere il tipo di cosa che potrebbe avere esperienza — senza però derivare da queste condizioni l'esistenza dell'esperienza stessa. In altri termini, il FEP descrive con precisione inedita la geometria dello spazio in cui il problema difficile si colloca, ma non lo occupa. Questa chiarificazione è filosoficamente preziosa: ci dice dove cercare, anche se non ci dice cosa troveremo.

Nei capitoli che seguono si procederà dunque in questo ordine. Si esporranno i fondamenti matematici del FEP, con derivazioni dettagliate che permettano al lettore di apprezzare la struttura tecnica del principio senza doverla assumere per fede. Si esaminerà poi la teoria dell'inferenza attiva come estensione del FEP all'azione, e si discuterà il concetto di confine di Markov come fondamento ontologico della distinzione soggetto/ambiente. Si passerà quindi alle implicazioni filosofiche, affrontando il problema difficile, la questione del sé come costrutto inferenziale, e la relazione tra il FEP e le principali posizioni nel dibattito sulla coscienza: il panpsichismo, l'illusionismo, l'enattivismo. Si concluderà con un bilancio critico e con alcune indicazioni su dove, a giudizio di chi scrive, una teoria completa della mente dovrebbe spingersi oltre il FEP.

V. Una nota sul metodo

Prima di addentrarci nel dettaglio tecnico, è opportuna una precisazione metodologica. Questo saggio non intende essere una rassegna neutrale della letteratura: prende posizioni,

argomenta a favore di tesi e contro altre, e non esita a muovere obiezioni a Friston dove le ritiene fondate. Al tempo stesso, si sforza di restare ancorato a quanto si conosce con sufficiente sicurezza, segnalando esplicitamente dove il terreno si fa speculativo. Il confine tra la neuroscienza computazionale come disciplina empirica e il FEP come programma metafisico è una frontiera che il testo attraverserà più volte, e ogni attraversamento sarà marcato.

Si adotta qui un'epistemologia moderatamente naturalistica: la filosofia non può procedere ignorando i risultati delle scienze, ma la scienza non esaurisce il dominio delle domande filosoficamente significative. Il FEP è un oggetto di studio che si presta a questo tipo di analisi binoculare con particolare fecondità, perché la sua ambizione teorica eccede sistematicamente i suoi risultati empirici — e questa eccedenza è proprio il luogo dove la filosofia ha qualcosa di essenziale da dire.

I Fondamenti Matematici

I. L'inferenza bayesiana come punto di partenza

Ogni esposizione rigorosa del FEP deve prendere le mosse dall'inferenza bayesiana, non perché il FEP si riduca ad essa, ma perché la struttura matematica del FEP è costruita come un'estensione e un'approssimazione dell'inferenza bayesiana esatta a sistemi per i quali tale inferenza sarebbe computazionalmente intrattabile.

Si consideri un agente — per ora semplicemente un sistema fisico con stati interni osservabili — che riceve segnali sensoriali denotati con $\mathbf{o}$ (dall'inglese *observations*). Questi segnali sono generati da cause nel mondo che l'agente non osserva direttamente: le variabili latenti o stati nascosti del mondo, denotate con ψ (psi). La relazione tra stati nascosti e osservazioni è codificata in un modello generativo, ossia nella distribuzione di probabilità congiunta:

$$p(o, \psi) = p(o \mid \psi) \cdot p(\psi)$$

dove $p(o \mid \psi)$ è la verosimiglianza — la probabilità di osservare $\mathbf{o}$ dato che il mondo si trova nello stato ψ — e $p(\psi)$ è la distribuzione a priori sugli stati del mondo, che codifica le aspettative dell'agente prima di ricevere qualsiasi osservazione.

Il problema inferenziale che l'agente deve risolvere è il seguente: date le osservazioni $\mathbf{o}$, qual è la distribuzione a posteriori $p(\psi \mid o)$ sugli stati nascosti? Per il teorema di Bayes:

$$p(\psi \mid o) = p(o \mid \psi) \cdot p(\psi) \,/\, p(o)$$

dove il denominatore, la cosiddetta evidenza del modello o marginalizzazione della verosimiglianza, è:

$$p(o) = \int p(o \mid \psi) \cdot p(\psi)\, d\psi$$

È questo integrale il nodo gordiano di ogni teoria bayesiana del cervello. Nella stragrande maggioranza dei modelli cognitivamente plausibili, l'integrale sulla destra è intrattabile: lo spazio degli stati nascosti è ad altissima dimensionalità, le distribuzioni coinvolte sono non gaussiane, e il calcolo esatto richiederebbe risorse computazionali esponenzialmente crescenti con la complessità del modello. Il cervello non può fare inferenza bayesiana esatta — e tuttavia si comporta, in molti contesti, come se lo facesse. Come riconciliare questi due fatti?

II. L'inferenza variazionale e l'energia libera

La risposta è l'inferenza variazionale: invece di calcolare la vera distribuzione a posteriori $p(\psi \mid o)$, si cerca la migliore approssimazione a essa all'interno di una famiglia di distribuzioni trattabili $q(\psi ; \lambda)$, parametrizzata da parametri variazionali λ. La qualità dell'approssimazione è misurata dalla divergenza di Kullback-Leibler tra q e la vera posteriore:

$$KL[q(\psi ; \lambda) \;//\; p(\psi \mid o)] = \int q(\psi ; \lambda) \cdot \ln[q(\psi ; \lambda) / p(\psi \mid o)]\, d\psi$$

Questa quantità è sempre non-negativa (per la disuguaglianza di Jensen) ed è nulla se e solo se q coincide con la vera posteriore. Minimizzare la KL rispetto a λ è la strategia di inferenza variazionale.

Il problema è che per minimizzare questa KL è necessario conoscere $p(\psi \mid o)$, che è precisamente ciò che stiamo cercando. Si introduce allora l'energia libera variazionale **F**, che si ottiene riscrivendo la KL in modo da eliminare la dipendenza dalla posteriore vera. Si parte dalla definizione:

$$KL[q(\psi) \;//\; p(\psi \mid o)] = \int q(\psi) \cdot \ln[q(\psi) / p(\psi \mid o)]\, d\psi$$

Sostituendo $p(\psi \mid o) = p(o, \psi) / p(o)$:

$$KL[q(\psi) \;//\; p(\psi \mid o)] = \int q(\psi) \cdot \ln[q(\psi) \cdot p(o) / p(o, \psi)]\, d\psi$$

$$= \int q(\psi) \cdot [\ln q(\psi) - \ln p(o, \psi) + \ln p(o)]\, d\psi$$

$$= \int q(\psi) \cdot [\ln q(\psi) - \ln p(o, \psi)]\, d\psi + \ln p(o)$$

Si definisce allora l'energia libera variazionale come:

$$F[q, o] \equiv \int q(\psi) \cdot [\ln q(\psi) - \ln p(o, \psi)] \, d\psi$$

$$= \mathbb{E}_q[\ln q(\psi) - \ln p(o, \psi)]$$

da cui segue immediatamente:

$$KL[q(\psi) \;//\; p(\psi \mid o)] = F[q, o] - \ln p(o)$$

Poiché la KL è sempre ≥ 0:

$$F[q, o] \geq -\ln p(o)$$

Questa disuguaglianza è di importanza capitale. Il termine -ln p(o) è la sorpresa (*surprisal*) o autocodifica (*self-information*) dell'osservazione **o** sotto il modello generativo: misura quanto l'osservazione è inattesa dal punto di vista dell'agente. Poiché F è un limite superiore della sorpresa, minimizzare F equivale a minimizzare un upper bound sulla sorpresa — ossia a rendere le osservazioni il più possibile attese, compatibilmente con le capacità computazionali del sistema.

L'energia libera F dipende sia da q (le credenze dell'agente sugli stati del mondo) sia da **o** (le osservazioni). L'inferenza bayesiana approssimata consiste nel minimizzare F rispetto a q, per fissate osservazioni **o**. Ciò che Friston aggiunge è la possibilità di minimizzare F anche rispetto alle azioni — cambiando **o** campionando osservazioni più compatibili con il modello generativo. Questa è l'inferenza attiva, di cui si tratterà in seguito.

III. Le decomposizioni dell'energia libera

La ricchezza concettuale del FEP emerge dalle molteplici decomposizioni equivalenti dell'energia libera, ciascuna delle quali illumina un aspetto diverso della struttura inferenziale.

Prima decomposizione: precisione e complessità

Si può scrivere:

$$\ln p(o, \psi) = \ln p(o \mid \psi) + \ln p(\psi)$$

Sostituendo nella definizione di F:

$$F = \mathbb{E}_q[\ln q(\psi)] - \mathbb{E}_q[\ln p(o \mid \psi)] - \mathbb{E}_q[\ln p(\psi)]$$

$$= \mathbb{E}_q[\ln q(\psi) - \ln p(\psi)] - \mathbb{E}_q[\ln p(o \mid \psi)]$$

$$= KL[q(\psi) \;//\; p(\psi)] - \mathbb{E}_q[\ln p(o \mid \psi)]$$

Il primo termine è la **complessità**: misura quanto la distribuzione a posteriori approssimata $q(\psi)$ si discosta dalla distribuzione a priori $p(\psi)$. In fisica statistica, questa quantità è analoga all'entropia relativa; in teoria dell'informazione, corrisponde al costo in bit necessari per codificare la credenza aggiornata a partire dalla credenza a priori. Un sistema che minimizza F cercherà di mantenere la propria a posteriore vicina alla propria a priori — il che significa che aggiornerà le proprie credenze solo nella misura in cui l'evidenza lo richiede.

Il secondo termine è la **precisione** (o accuratezza): misura quanto il modello generativo rende conto delle osservazioni, mediato sulla distribuzione q. Un sistema che massimizza l'accuratezza cerca stati nascosti che spieghino bene i dati sensoriali.

La minimizzazione di F è dunque un compromesso tra due spinte opposte: minimizzare la complessità (restare vicino al priore) e massimizzare l'accuratezza (spiegare i dati). Questo compromesso è matematicamente identico al principio di Occam nell'inferenza bayesiana: il modello migliore è quello che spiega i dati con la minima complessità aggiuntiva.

F = Complessità - Precisione

$$F = KL[q(\psi) \;//\; p(\psi)] - \mathbb{E}_q[\ln p(o \mid \psi)]$$

Seconda decomposizione: energia ed entropia

In alternativa, espandendo il logaritmo del modello generativo:

$$F = -\mathbb{E}_q[\ln p(o, \psi)] - H[q(\psi)]$$

dove $H[q(\psi)] = -\int q(\psi) \ln q(\psi)\, d\psi$ è l'entropia di Shannon della distribuzione q. Questa scrittura rende perspicuo il legame con la meccanica statistica: il termine $-\mathbb{E}_q[\ln p(o, \psi)]$ è l'energia media del sistema (intendendo $p(o, \psi)$ come la misura di Gibbs del sistema fisico), mentre $H[q(\psi)]$ è l'entropia. L'energia libera variazionale è quindi formalmente analoga all'energia libera di Helmholtz della termodinamica:

$$F = \mathbb{E}[\text{Energia}] - T \cdot \text{Entropia}$$

dove la temperatura T è implicitamente unitaria (o assorbita nella definizione di energia). Questa connessione non è meramente analogica: Friston l'ha esplorata sistematicamente per suggerire che i sistemi biologici auto-organizzati occupano uno spazio tra il minimo di energia e il massimo di entropia — uno spazio che è precisamente quello occupato dai sistemi termodinamici che si trovano lontano dall'equilibrio ma non al caos.

Terza decomposizione: sorpresa e informazione mutua

Una terza scrittura equivalente è:

$$F = -\ln p(o) + KL[q(\psi) \;//\; p(\psi \mid o)]$$

$$= \text{Sorpresa} + \text{KL residuo}$$

Questa decomposizione chiarisce che l'energia libera maggiora la sorpresa di un ammontare pari alla KL tra la densità di riconoscimento q e la vera a posteriori. Quando q coincide esattamente con la vera a posteriori, l'energia libera collassa sulla sorpresa e l'inferenza variazionale diventa esatta. La minimizzazione di F rispetto a q (per **o** fisso) riduce il KL residuo, avvicinando q alla vera posteriore; la minimizzazione di F rispetto ad **o** (ossia tramite l'azione) riduce direttamente la sorpresa.

Queste tre decomposizioni non sono ridondanti: ciascuna apre una finestra su aspetti diversi del processo inferenziale e suggerisce diverse implementazioni neurali. Il predictive coding, come si vedrà, è naturalmente derivato dalla prima decomposizione; la connessione con la termodinamica dei sistemi biologici emerge dalla seconda; la struttura della riduzione della sorpresa tramite l'azione è più chiara nella terza.

IV. Il modello generativo gerarchico

Nella teoria del FEP applicata al cervello, il modello generativo non è una distribuzione piatta su stati nascosti scalari, ma una struttura gerarchica che riflette la struttura causale multilivello del mondo. Si denotano con $\psi^{(1)}, \psi^{(2)}, ..., \psi^{(L)}$ gli stati nascosti a livelli successivamente più astratti della gerarchia, con o le osservazioni al livello inferiore (input sensoriale). Il modello generativo gerarchico ha la forma:

$$p(o, \psi^{(1)}, ..., \psi^{(L)}) = p(o \mid \psi^{(1)}) \cdot \prod_{l=1}^{L-1} p(\psi^{(l)} \mid \psi^{(l+1)}) \cdot p(\psi^{(L)})$$

Ogni livello della gerarchia funge da modello degli stati del livello inferiore: $\psi^{(2)}$ è la causa di $\psi^{(1)}$, che a sua volta è la causa delle osservazioni **o**. In una versione continua e dinamica, gli stati nascosti evolvono nel tempo secondo equazioni differenziali stocastiche:

$$d\psi^{(l)}/dt = f^{(l)}(\psi^{(l)}, \psi^{(l+1)}) + \omega^{(l)}$$

dove $f^{(l)}$ è una funzione (potenzialmente non lineare) che descrive la dinamica al livello l condizionata agli stati del livello superiore, e $\omega^{(l)}$ è un termine di rumore gaussiano con precisione (inverso della varianza) $\Pi^{(l)}$. Le osservazioni sono generate dal livello più basso:

$$o = g(\psi^{(1)}) + \varepsilon$$

dove g è una funzione di osservazione e ε è un rumore sensoriale con precisione $\Pi^{(o)}$.

Questa struttura gerarchica è cruciale per due ragioni. In primo luogo, permette al modello di rappresentare le regolarità del mondo a scale temporali e spaziali diverse: le cause veloci e locali al livello più basso, le cause lente e globali ai livelli più alti. In secondo luogo, fornisce un'architettura naturale per il predictive coding, come si mostrerà tra breve.

V. L'approssimazione di Laplace e la gaussianizzazione

Per rendere l'inferenza variazionale computazionalmente trattabile in un modello gerarchico come quello descritto, Friston adotta sistematicamente l'approssimazione di Laplace: la distribuzione a posteriori approssimata $q(\psi)$ viene assunta gaussiana, centrata sulla moda della posteriore vera. Questa è una scelta radicale che sacrifica la generalità per la trattabilità, e le sue implicazioni meritano attenzione critica.

Formalmente, se si denota con μ il vettore delle medie (le credenze medie sugli stati nascosti) e con Σ la matrice di covarianza, allora:

$$q(\psi) = \mathcal{N}(\psi\,;\,\mu,\,\Sigma)$$

e l'energia libera, sotto questa assunzione, diventa:

$$F_{\text{Laplace}} = -\ln p(o,\mu) - \tfrac{1}{2}\ln\det(2\pi\Sigma) + \tfrac{1}{2}\,\mathrm{tr}\left(\Sigma\cdot\partial^2[-\ln p(o,\psi)]/\partial\psi^2\,|_{\psi=\mu}\right)$$

Nel caso in cui la precisione della posteriore sia elevata (bassa varianza), l'approssimazione di Laplace diventa sempre più accurata. L'aggiornamento delle credenze medie μ che minimizza F_{Laplace} segue un'equazione di punto fisso della forma:

$$\dot{\mu} = -\partial F/\partial\mu = \partial[\ln p(o,\mu)]/\partial\mu$$

ossia un processo di discesa del gradiente sull'energia libera nello spazio dei valori medi. Questo aggiornamento è la base matematica per il predictive coding.

Con l'approssimazione di Laplace e il modello gerarchico, l'energia libera si semplifica considerevolmente. Denotando con $\varepsilon^{(l)}$ gli errori di predizione al livello l:

$$\varepsilon^{(o)} = o - g(\mu^{(1)}) \qquad \text{[errore sensoriale]}$$
$$\varepsilon^{(l)} = \mu^{(l)} - f^{(l)}(\mu^{(l+1)}) \quad \text{[errore di predizione al livello } l\text{]}$$

l'energia libera assume la forma quadratica:

$$F \approx \tfrac{1}{2}\left[\,(\varepsilon^{(o)})^{\mathsf{T}}\,\Pi^{(o)}\,\varepsilon^{(o)} + \sum_{l=1}^{L}(\varepsilon^{(l)})^{\mathsf{T}}\,\Pi^{(l)}\,\varepsilon^{(l)}\,\right] + \text{const}$$

dove le matrici di precisione $\Pi^{(l)}$ pesano gli errori di predizione ai diversi livelli. L'energia libera è dunque una somma pesata di errori di predizione al quadrato, dove i pesi sono le precisioni — gli inversi delle varianze. La minimizzazione di F equivale pertanto a minimizzare la somma pesata degli errori di predizione in tutta la gerarchia.

VI. Il predictive coding come implementazione neurale

La derivazione precedente conduce direttamente all'architettura del predictive coding. Definiamo le variabili:

$$\xi^{(l)} \equiv \Pi^{(l)} \cdot \varepsilon^{(l)}$$

come le unità di errore di predizione ponderate dalla precisione al livello l. Le equazioni di aggiornamento che minimizzano F rispetto alle medie $\mu^{(l)}$ sono, in forma dinamica:

$$\tau \cdot d\mu^{(l)}/dt = -\xi^{(l)} + (\partial f^{(l)}/\partial \mu^{(l)})^{T} \cdot \xi^{(l+1)}$$

dove τ è una costante di tempo. Il primo termine ($-\xi^{(l)}$) spinge $\mu^{(l)}$ verso valori che riducono l'errore di predizione al livello l; il secondo termine ($(\partial f/\partial \mu)^{T} \cdot \xi^{(l+1)}$) trasmette il segnale di errore del livello superiore verso il basso. Gli errori di predizione $\xi^{(l)}$ sono a loro volta aggiornati in base alla differenza tra le predizioni generate dal livello superiore e i valori effettivi:

$$\xi^{(l)} = \Pi^{(l)} \cdot [\mu^{(l)} - f^{(l)}(\mu^{(l+1)})]$$

Queste equazioni descrivono un sistema dinamico in cui le unità di stato $\mu^{(l)}$ e le unità di errore $\xi^{(l)}$ interagiscono reciprocamente: le prime generano predizioni che vengono confrontate con i segnali provenienti dal basso, le seconde propagano verso l'alto le discrepanze. Questa è precisamente l'architettura del predictive coding: un processo in cui la gerarchia corticale è interpretata come un sistema di soppressione iterativa degli errori di predizione.

La corrispondenza anatomica proposta è la seguente. Le connessioni discendenti (*top-down* o *feedback*) dalla corteccia di ordine superiore alle aree di ordine inferiore trasportano le predizioni generate dal modello interno; le connessioni ascendenti (*bottom-up* o *feedforward*) trasportano gli errori di predizione. Le cellule piramidali superficiali, che proiettano in avanti verso aree di ordine superiore, sarebbero le unità di errore ξ; le cellule piramidali profonde e le cellule stellate, che ricevono input feedforward e generano predizioni, sarebbero le unità di stato μ. Questa mappatura è compatibile con la laminare struttura della corteccia e con i pattern di connettività osservati, anche se la verifica empirica rigorosa rimane un programma di ricerca aperto.

VII. La precisione e il ruolo dell'attenzione

Una delle conseguenze più eleganti del formalismo variazionale è la naturalizzazione dell'attenzione. Nel contesto del FEP, la precisione $\Pi^{(l)}$ al livello l non è un parametro fisso: è essa stessa una variabile che il sistema ottimizza. La minimizzazione di F rispetto alle precisioni conduce a una stima della precisione segnale-a-rumore ai diversi livelli della gerarchia.

Formalmente, la stima ottimale della precisione $\hat{\Pi}^{(l)}$ è:

$$\hat{\Pi}^{(l)} = (\mathbb{E}_q[(\varepsilon^{(l)})(\varepsilon^{(l)})^T])^{-1}$$

ossia l'inverso della covarianza degli errori di predizione. Questo ha un'interpretazione immediata: la precisione stimata è alta quando gli errori di predizione sono piccoli e consistenti (alta certezza), e bassa quando sono grandi e variabili (alta incertezza). Il sistema assegna maggiore peso ai livelli o alle modalità sensoriali in cui gli errori di predizione sono più informativi.

L'attenzione, in questa prospettiva, è il processo di regolazione dinamica delle precisioni: dirigere l'attenzione verso uno stimolo equivale ad aumentare la precisione delle unità di errore che codificano quell'aspetto dell'input, amplificando il loro segnale nella competizione tra ipotesi rivali. In termini neurali, la trasmissione neuromodulatoriale — in particolare quella dopaminergica e colinergica — è stata proposta come substrato di questa regolazione della precisione, con la dopamina come segnale di precisione del priore e l'acetilcolina come segnale di precisione della verosimiglianza. Queste proposte trovano un certo supporto nella psicofarmacologia delle psicosi, in cui le alterazioni della neurotrasmissione dopaminergica sembrano produrre distorsioni nella ponderazione relativa dei segnali di errore di predizione.

VIII. Limiti matematici dell'approssimazione gaussiana

Prima di procedere alle implicazioni filosofiche, è necessario segnalare con franchezza i limiti matematici dell'architettura descritta, perché sono rilevanti non solo per la validità empirica del FEP ma anche per le sue pretese filosofiche.

L'approssimazione di Laplace funziona bene quando la distribuzione a posteriori è unimodale e ben approssimata da una gaussiana in un intorno della moda. Quando la posteriore è multimodale — come avviene in scenari di ambiguità percettiva o di inferenza su cause rivali — l'approssimazione gaussiana è gravemente inadeguata. In tali situazioni, un sistema che minimizza l'energia libera variacionale gaussiana può convergere su una sola delle ipotesi plausibili, mancando la struttura multimodale della vera posteriore.

Esistono estensioni del formalismo che affrontano parzialmente questo problema: le distribuzioni a miscela gaussiana (Gaussian Mixture Models), l'inferenza variacionale

strutturata, e i recenti sviluppi basati su reti neurali profonde che possono approssimare distribuzioni arbitrariamente complesse (come nelle reti di Helmholtz e nei modelli a variabili latenti profondi). Tuttavia, l'architettura di base del predictive coding gaussiano — che è quella per cui esiste la corrispondenza anatomica più diretta — è limitata al regime unimodale.

Questo limite è rilevante filosoficamente perché alcune delle pretese più ambiziose del FEP — ad esempio, la capacità di spiegare la coscienza come processo inferenziale — dipendono dall'assumere che il formalismo sia sufficientemente generale da coprire l'intera fenomenologia mentale. Se l'approssimazione gaussiana non riesce a catturare la struttura di incertezza nelle decisioni percettive quotidiane, è difficile immaginare che possa catturare la ricchezza dell'esperienza soggettiva.

Un secondo limite riguarda l'identificabilità del modello generativo. Il FEP assume che il sistema neurobiologico abbia un modello generativo $p(o, \psi)$, ma non specifica da dove questo modello provenga — come sia stato appreso, quali siano i suoi parametri, e come si modifichi nel corso dell'esperienza. L'apprendimento nel contesto del FEP è trattato come minimizzazione dell'energia libera rispetto ai parametri del modello (oltre che alle variabili latenti), e la plasticità sinaptica è proposta come substrato neurale di questo processo. Ma il problema di come il cervello impari modelli generativi corretti del mondo a partire dall'esperienza sensoriale è uno dei problemi più difficili della neuroscienza computazionale, e il FEP da solo non lo risolve: lo enuncia in un linguaggio formale, spostando il problema a un livello più astratto.

Queste osservazioni non sono intese a sminuire il valore del formalismo: sono intese a delimitare con precisione il dominio in cui le sue pretese sono giustificate. E questa delimitazione è essenziale per affrontare, con la necessaria onestà intellettuale, le implicazioni filosofiche del FEP — che sono il tema della sezione successiva.

Inferenza Attiva e il Confine di Markov come Ontologia del Sé

I. Dall'inferenza passiva all'inferenza attiva

Fino a questo punto si è trattato il problema dell'inferenza percettiva come se l'agente fosse un ricevitore passivo di segnali sensoriali: il mondo genera osservazioni, il cervello le elabora inferendo le cause nascoste. Questa asimmetria, tuttavia, è una finzione utile ma ontologicamente fuorviante. Gli organismi biologici non osservano il mondo: vi agiscono, e la loro azione modifica il flusso sensoriale che ricevono. È questa circolarità tra azione e percezione — l'arco sensorimorale che Merleau-Ponty aveva descritto fenomenologicamente e che Gibson aveva concettualizzato nell'ecologia della percezione — che il FEP deve

integrare nel suo formalismo se vuole pretendere di essere una teoria della mente e non soltanto una teoria della percezione passiva.

L'estensione del FEP all'azione prende il nome di inferenza attiva (*active inference*), e la sua struttura matematica è sorprendentemente elegante. L'intuizione fondamentale è che l'agente può minimizzare l'energia libera non soltanto modificando le proprie credenze sugli stati nascosti (aggiornamento percettivo) ma anche modificando le osservazioni stesse, agendo sul mondo in modo da renderle più compatibili con le proprie aspettative. In termini formali, sia **a** un vettore di variabili d'azione che parametrizzano come l'agente modifica il proprio input sensoriale. Le osservazioni **o** dipendono ora sia dagli stati nascosti del mondo sia dalle azioni dell'agente:

$$o = g(\psi, a) + \varepsilon$$

L'energia libera variazionale $F[q, o(a)]$ è ora una funzione anche di **a** attraverso la sua dipendenza dalle osservazioni. Il gradiente di F rispetto ad **a** è:

$$\partial F/\partial a = \partial F/\partial o \cdot \partial o/\partial a = -\hat{\xi}(o) \cdot (\partial g/\partial a)^{T}$$

ossia l'azione è guidata dagli errori di predizione sensoriali $\hat{\xi}(o)$ pesati dalla Jacobiana della funzione di osservazione rispetto all'azione. Un'interpretazione immediata: l'azione tende a ridurre la discrepanza tra ciò che il sistema si aspetta di sentire e ciò che effettivamente sente — essa realizza le predizioni piuttosto che aggiornarle.

Questo ha una conseguenza concettuale di vasta portata: nella prospettiva dell'inferenza attiva, l'azione non è l'output di un sistema di pianificazione separato che riceve input dalla percezione; è il complemento della percezione all'interno di un unico processo di minimizzazione. La distinzione tra percezione e azione è funzionale, non ontologica: entrambe minimizzano l'energia libera, ma percorrendo direzioni diverse nello spazio delle variabili che il sistema può modificare.

II. Il problema dell'azione come realizzazione delle profezie

L'inferenza attiva introduce tuttavia un paradosso apparente che merita attenzione. Se l'azione riduce gli errori di predizione facendo sì che le osservazioni corrispondano alle aspettative, sembra che il sistema sia condannato a una sorta di solipsismo predittivo: si muove nel mondo per confermare ciò che già si aspetta, trasformando il mondo in uno specchio delle proprie credenze. Come può un tale sistema apprendere qualcosa di genuinamente nuovo? Come può correggere modelli sbagliati?

La risposta sta nella struttura gerarchica del modello generativo e nella distinzione tra priori a diversi livelli. Al livello più basso, le predizioni sensoriali immediate vengono realizzate dall'azione; ma al livello più alto, le aspettative sulle cause del mondo devono confrontarsi

con le conseguenze delle proprie azioni — e queste conseguenze possono disconfirmare le aspettative di livello superiore anche mentre le azioni le confermano a livello inferiore.

Più precisamente, Friston distingue tra aspettative sulle traiettorie sensoriali (*proprioceptive predictions*) e aspettative sugli stati del mondo (*exteroceptive model*). Le prime sono realizzate direttamente dagli archi riflessi, che fungono da meccanismi di controllo che portano il sistema sensoriale a corrispondere alla predizione muscolare. Le seconde sono aggiornate passivamente dall'evidenza sensoriale che emerge dalle azioni. L'apprendimento avviene quando le conseguenze delle azioni modificano le credenze sui livelli superiori della gerarchia, inducendo aggiornamenti nel modello del mondo che non erano stati anticipati dalle predizioni di livello inferiore.

In termini matematici, sia μ^(alto) il vettore delle credenze al livello più alto della gerarchia. L'equazione di aggiornamento per μ^(alto) durante l'azione è:

$$\tau \cdot d\mu^{(alto)}/dt = -(\partial F/\partial \mu^{(alto)})|_{o = o(a(t))}$$

dove o(a(t)) sono le osservazioni che risultano dall'azione attuale. Se queste osservazioni differiscono sistematicamente dalle predizioni generate da μ^(alto), l'aggiornamento spinge μ^(alto) verso valori che le spieghino meglio — anche se, nel frattempo, l'azione stava minimizzando gli errori di predizione di livello inferiore. La gerarchia funziona dunque come un sistema di aggiornamento a scala multipla: veloce alla base, lento all'apice.

III. Le politiche come distribuzioni su traiettorie

La formulazione dell'inferenza attiva che si è presentata fin qui è cinematica: descrive come l'azione istantanea minimizza l'energia libera. Ma gli organismi non scelgono azioni istantanee isolate: scelgono *politiche* — sequenze di azioni estese nel tempo che portano a stati futuri. L'estensione del FEP alla pianificazione temporale richiede un'ulteriore elaborazione.

Friston introduce la nozione di *expected free energy* (energia libera attesa) $G(\pi)$ associata a una politica π, che si distingue dall'energia libera variazionale F per il fatto di essere calcolata rispetto a stati futuri non ancora osservati. Formalmente, per una politica π che genera una sequenza di azioni $a_1, a_2, ..., a_T$:

$$G(\pi) = \mathbb{E}_{q(o_\tau, \psi_\tau \mid \pi)}[\ \ln q(\psi_\tau \mid \pi) - \ln p(o_\tau, \psi_\tau)\ ]$$

dove la media è calcolata rispetto alla distribuzione congiunta futura su osservazioni e stati nascosti sotto la politica π. Si può mostrare che questa quantità si decompone in due termini:

$$G(\pi) = -\mathbb{E}_{q(o_\tau \mid \pi)}[\ \ln p(o_\tau)\ ] - \mathbb{E}_{q(\psi_\tau \mid \pi)}[\ H[p(o_\tau \mid \psi_\tau)]\ ]$$

$$\approx \text{Rischio (KL rispetto alle preferenze)} + \text{Ambiguità (incertezza epistemica)}$$

Il primo termine misura la discrepanza tra le osservazioni future attese sotto π e le osservazioni preferite dall'agente — le osservazioni che il suo modello generativo assegna alta probabilità a priori. Minimizzare questo termine orienta l'agente verso stati preferiti, ossia verso stati con bassa sorpresa attesa. Il secondo termine misura l'ambiguità: la varianza attesa nella mappatura da stati nascosti a osservazioni. Minimizzare l'ambiguità spinge l'agente verso azioni che risolvono l'incertezza sul mondo — verso comportamenti esplorativi ed epistemici.

La distribuzione sulle politiche è quindi:

$$q(\pi) \propto \exp(-G(\pi))$$

e la politica ottima è quella che minimizza l'energia libera attesa. Questo formalismo incorpora naturalmente sia la motivazione strumentale (ridurre il rischio, avvicinarsi agli stati preferiti) sia la motivazione epistemica (ridurre l'ambiguità, raccogliere informazioni). L'unificazione di motivazione strumentale e motivazione epistemica in un unico principio è uno dei contributi concettuali più significativi dell'inferenza attiva, perché dissolve la dicotomia tra esplorazione e sfruttamento che aveva dominato la letteratura sul rinforzo.

Occorre notare che l'energia libera attesa G(π) non è la stessa cosa dell'energia libera variazionale F: la prima è un funzionale sulle politiche che richiede di integrare su stati futuri non ancora osservati; la seconda è un funzionale sulle credenze attuali rispetto alle osservazioni presenti. La relazione tra le due è quella tra pianificazione e percezione: la minimizzazione di F aggiorna le credenze date le osservazioni attuali; la minimizzazione di G seleziona le politiche che minimizzeranno la sorpresa attesa nel futuro.

IV. Il confine di Markov: la matematica dell'individualità

Il contributo forse più filosoficamente carico del programma di Friston è l'analisi del concetto di *confine di Markov* (*Markov blanket*), introdotto originariamente da Judea Pearl nell'ambito delle reti bayesiane e radicalmente reinterpretato da Friston come strumento ontologico per definire l'individualità di un sistema.

Nella teoria delle reti bayesiane, il confine di Markov di un nodo X è l'insieme minimale di nodi che, se condizionato, rende X statisticamente indipendente da tutti gli altri nodi della rete. Formalmente, in una rete bayesiana con nodi {X, Y, Z, ...}, il confine di Markov b(X) di X è tale che:

$$p(X \mid b(X), \text{tutti gli altri nodi}) = p(X \mid b(X))$$

ossia X è condizionalmente indipendente dal resto della rete, dato il suo confine. Il confine di Markov di un nodo in una rete bayesiana è composto dai suoi genitori, dai suoi figli e dai co-genitori dei suoi figli — l'insieme delle variabili che mediano ogni influenza causale tra X e il resto del sistema.

Friston estende questo concetto ai sistemi dinamici stocastici. Si consideri un sistema con un gran numero di variabili che interagiscono secondo un processo di Langevin:

$dz/dt = f(z) + \omega$

dove $z = (\eta, s, a, \mu)$ è il vettore di stato totale del sistema, ω è un termine di rumore gaussiano, e la funzione di deriva f determina le relazioni causali tra le variabili. Si definisce una partizione delle variabili in quattro sottoinsiemi:

η : stati esterni (del mondo, al di fuori del sistema)
s : stati sensoriali (che mediano l'influenza di η su μ)
a : stati di azione (che mediano l'influenza di μ su η)
μ : stati interni (del sistema, al di fuori del confine)

Il confine di Markov è dato dall'unione degli stati sensoriali e degli stati di azione: $b = (s, a)$. Le proprietà che questa partizione deve soddisfare per costituire un confine di Markov autentico sono:

f_η non dipende da μ (gli stati esterni non sono causalmente influenzati dagli stati interni direttamente)
f_μ non dipende da η (gli stati interni non sono causalmente influenzati dagli stati esterni direttamente)

In altre parole, il confine di Markov implica che l'unica via di comunicazione causale tra l'interno e l'esterno del sistema passa attraverso il confine stesso — gli stati sensoriali trasmettono influenze dall'esterno verso l'interno, gli stati di azione trasmettono influenze dall'interno verso l'esterno, ma non esiste accoppiamento diretto tra stati interni e stati esterni.

Formalmente, la densità di probabilità stazionaria del sistema p(z) si fattorizza secondo la struttura di dipendenza determinata dal confine:

$p(\eta, \mu \mid s, a) = p(\eta \mid s, a) \cdot p(\mu \mid s, a)$

ossia, condizionati al confine di Markov, gli stati interni e gli stati esterni sono statisticamente indipendenti.

V. Dal confine di Markov all'auto-organizzazione

La rilevanza del confine di Markov per il FEP emerge da un risultato di Friston che merita una derivazione attenta. Si consideri la distribuzione stazionaria del processo di Langevin:

$p(z) = p(\eta, s, a, \mu)$

Per la struttura del confine di Markov, questa si può scrivere come:

$$p(\eta, s, a, \mu) = p(\eta \mid b) \cdot p(\mu \mid b) \cdot p(b)$$

dove $b = (s, a)$. Ora si consideri la distribuzione marginale degli stati interni dato il confine:

$$p(\mu \mid b)$$

e si definisca la distribuzione *di riconoscimento* come la distribuzione degli stati interni condizionata al confine: $q_b(\eta) \equiv p(\eta \mid \mu, b)$. Per la proprietà del confine di Markov, $p(\eta \mid \mu, b) = p(\eta \mid b)$, quindi la distribuzione di riconoscimento non dipende direttamente da μ — o, più precisamente, dipende da μ attraverso la sua influenza sul confine b.

Si può ora mostrare che la densità di probabilità stazionaria degli stati interni può essere scritta come:

$$p(\mu \mid b) \propto \exp(-F(\mu, b))$$

dove $F(\mu, b)$ è una misura di energia libera degli stati interni dato il confine. Questo risultato, nella sua versione rigorosa, richiede alcune assunzioni sulla struttura del rumore e sulla stazionarietà del processo, ma il suo contenuto è intuitivamente chiaro: la distribuzione degli stati interni di un sistema con un confine di Markov è equivalente alla distribuzione di Boltzmann di un sistema che minimizza una funzione di energia libera.

Il corollario di importanza filosofica è il seguente: qualsiasi sistema fisico dotato di un confine di Markov stabile può essere descritto *come se* i suoi stati interni stessero compiendo inferenza bayesiana sugli stati esterni mediata dal confine. Questo non significa che il sistema *davvero* compia inferenza — non implica la presenza di stati rappresentazionali, né tantomeno di esperienza soggettiva. Significa soltanto che il processo fisico ammette una descrizione nel vocabolario dell'inferenza. È una questione di equivalenza formale, non di riduzione ontologica.

VI. La gerarchia dei confini e l'emergenza del sé

Una delle idee più ardite di Friston è che i confini di Markov siano annidabili in una struttura gerarchica, e che questa struttura annidiata possa fornire un modello formale dell'emergenza del sé come costrutto inferenziale.

A livello cellulare, ogni cellula ha un confine di Markov: la membrana cellulare, con i suoi recettori (stati sensoriali) e i suoi canali ionici e proteine di secrezione (stati di azione), separa il metabolismo interno (stati interni) dall'ambiente extracellulare (stati esterni). L'organismo multicellulare può essere visto come un sistema i cui elementi hanno confini di Markov locali

e il cui collettivo ha un confine di Markov globale — la pelle, l'interfaccia sensorimotoria con il mondo.

Al livello neurologico, il confine di Markov dell'agente cognitivo è l'interfaccia tra il sistema nervoso (stati interni) e il corpo/mondo (stati sensoriali e di azione). Il modello generativo che il sistema nervoso mantiene — le credenze sugli stati nascosti del mondo — è una rappresentazione probabilistica del sistema esterno al confine. Quando questo modello generativo include una rappresentazione del confine stesso — quando il cervello modella la propria interfaccia con il mondo — emerge qualcosa che si può chiamare *modello del sé*: una rappresentazione probabilistica delle proprietà del sistema che possiede il confine.

Il sé, in questa prospettiva, non è una sostanza né un'entità: è un'ipotesi inferenziale. È la migliore spiegazione che il sistema nervoso trova per il pattern di correlazioni tra i propri stati sensoriali e i propri stati di azione — un'ipotesi sulla causa di queste correlazioni che risulta essere, in modo autoriflessivo, il sistema stesso. Karl Friston descrive questo come *autoevidenza* (*self-evidencing*): il sistema fa inferenza su se stesso attraverso le sue azioni nel mondo, e il sé è il contenuto di questa inferenza su se stessa.

Formalmente, si può scrivere la distribuzione a posteriori del sistema su se stesso come:

$$p(\text{sé} \mid b) \propto p(b \mid \text{sé}) \cdot p(\text{sé})$$

dove $p(b \mid \text{sé})$ è la verosimiglianza del confine dato il modello del sé, e $p(\text{sé})$ è il priore sul tipo di entità che il sistema si aspetta di essere. La minimizzazione dell'energia libera aggiorna sia il modello del mondo sia il modello del sé: il sé emerge come parte del modello generativo necessaria per spiegare il pattern sensorimotorio del sistema.

VII. Problemi ontologici del confine di Markov

La proposta del confine di Markov come fondamento ontologico dell'individualità è concettualmente audace, ma solleva problemi seri che vanno affrontati con onestà.

Il primo problema è quello dell'arbitrarietà della partizione. La definizione formale di confine di Markov richiede una partizione delle variabili del sistema in interne, esterne, sensoriali e di azione. Ma questa partizione non è una proprietà oggettiva del sistema fisico: dipende dalla scelta di quali variabili si includono nella descrizione e da come si le raggruppa. Per un sistema fisico con molti gradi di libertà, esistono molteplici partizioni compatibili con la struttura di dipendenza condizionale, e diverse partizioni danno confini di Markov diversi. La domanda «qual è il vero confine di Markov di un organismo?» non ha una risposta univoca determinata dalla fisica del sistema.

Questo non invalida il concetto, ma ne limita la portata ontologica. Il confine di Markov è un concetto prospettivo: è un confine relativo a una descrizione, non relativo alla cosa in sé. Un organismo non *ha* un confine di Markov come ha un cuore o un fegato: ha un confine di Markov relativo a un certo livello di descrizione e a una certa scelta di variabili rilevanti.

Questa prospettività è compatibile con il realismo scientifico, ma è incompatibile con la pretesa di derivare l'ontologia del sé direttamente dalla fisica statistica.

Il secondo problema riguarda la scala temporale della stazionarietà. L'identificazione del confine di Markov richiede la stazionarietà della distribuzione di probabilità del sistema — la sua convergenza a una distribuzione stazionaria stabile. Ma i sistemi biologici operano su molteplici scale temporali, alcune delle quali sono tutt'altro che stazionarie: un organismo che apprende, si sviluppa, si ammala e muore non ha una distribuzione stazionaria globale. L'applicazione del formalismo del confine di Markov ai sistemi biologici reali richiede quindi una localizzazione temporale — l'assunzione che il sistema sia quasi-stazionario su scale temporali di interesse — che limita la generalità del risultato.

Il terzo problema, e il più rilevante per le pretese filosofiche del FEP, è la distinzione tra avere un confine di Markov e avere stati rappresentazionali con contenuto intenzionale. Il risultato matematico mostra che qualsiasi sistema con un confine di Markov stabile può essere descritto nel vocabolario dell'inferenza. Ma questo non implica che il sistema abbia davvero credenze, desideri o esperienze: è una equivalenza tra descrizioni, non una riduzione ontologica. Un termostato ha qualcosa che assomiglia a un confine di Markov (la sonda termica è il suo stato sensoriale, il riscaldatore è il suo stato di azione, la temperatura interna è il suo stato interno), ma sarebbe bizzarro concludere che il termostato fa inferenza bayesiana e modella se stesso. La differenza tra il termostato e l'uomo non è di grado nell'avere confini di Markov: è — si potrebbe sostenere — di natura nell'essere il tipo di sistema per cui avere stati interni è qualcosa.

Ed è su questo punto preciso che il FEP, per quanto matematicamente sofisticato, incontra il suo limite filosofico fondamentale: un limite che costituisce l'oggetto della parte successiva.

Il Problema Difficile, le Posizioni Rivali e i Limiti del FEP

I. Il rientro del problema difficile

La struttura matematica del FEP, nella sua complessità e nella sua eleganza, rischia di produrre un effetto di abbaglio: l'impressione che il problema della coscienza sia stato risolto per accerchiamento, che il problema difficile si sia dissolto nel momento in cui si è mostrato che ogni aspetto della cognizione — percezione, azione, attenzione, sé — può essere derivato da un unico principio variazionale. Occorre resistere a questa impressione con decisione filosofica.

Il problema difficile, nella sua formulazione più netta, non riguarda quale meccanismo realizza la percezione, l'azione o il sé: riguarda perché qualsiasi meccanismo sia accompagnato

da esperienza soggettiva. La domanda non è «come funziona la percezione?» ma «perché la percezione è qualcosa che si sente?». Questi sono interrogativi logicamente distinti: la risposta alla prima domanda non implica necessariamente la risposta alla seconda. E il FEP, per quanto risponda con grande sofisticazione alla prima domanda, non tocca la seconda.

Per vedere questo con chiarezza, si consideri il seguente esperimento mentale. Si immagini un sistema fisico che implementa esattamente l'architettura del predictive coding gerarchico descritta nella Parte Seconda: la struttura neurale corretta, le equazioni differenziali giuste, le matrici di precisione opportunamente regolate, il confine di Markov ben definito. Questo sistema minimizza l'energia libera variazionale, realizza le politiche che minimizzano l'energia libera attesa, mantiene un modello generativo del mondo e di se stesso. Eppure, è logicamente concepibile che questo sistema svolga tutte queste funzioni senza che ci sia nulla che è come essere quel sistema — senza alcuna esperienza soggettiva. Questo è l'argomento dello zombie filosofico di Chalmers: la concevibilità logica di uno zombie funzionale suggerisce che la spiegazione funzionale non può, da sola, render conto dell'esperienza.

Si può obiettare che la concevibilità non implica la possibilità metafisica: siamo notoriamente cattivi giudici di ciò che è metafisicamente possibile. Kripke ha mostrato che alcune identità necessarie non sono a priori — acqua è necessariamente H_2O anche se questo non è conoscibile a priori. Forse l'esperienza soggettiva è necessariamente identica a certi processi di inferenza variazionale, anche se questo non è a priori evidente. Ma questa via — l'identità necessaria a posteriori — deve essere percorsa con molta cautela, perché l'asimmetria di concevibilità tra il caso acqua/H_2O e il caso esperienza/processo computazionale è notevole: sembra molto più difficile eliminare il residuo di concevibilità degli zombie nel secondo caso che nel primo.

II. L'explanatory gap nella geometria del FEP

Il contributo genuino del FEP al problema difficile non è di risolverlo ma di geometrizzarlo: di mostrare con precisione formale dove l'abisso esplicativo si apre. Questa chiarificazione ha un valore filosofico autonomo che non dipende dalla risoluzione del problema.

Si consideri la struttura del FEP come una mappa di tutto ciò che è spiegabile in termini di minimizzazione dell'energia libera. Su questa mappa si trovano: la struttura della percezione, le dinamiche dell'aggiornamento delle credenze, i meccanismi dell'azione volontaria e riflessa, la regolazione dell'attenzione, la formazione del sé come modello generativo autoriflessivo, persino alcune caratteristiche dell'emozione (intesa come precisione del priore interoceptivo, ossia come inferenza sugli stati corporei interni). Tutto questo è sul territorio del FEP.

Ma su questa mappa c'è una regione che rimane in bianco: la regione delle proprietà intrinseche degli stati interni. Il FEP descrive la *struttura relazionale* degli stati interni — come si relazionano agli input sensoriali, agli output di azione, agli altri stati interni — ma non dice nulla sulle loro proprietà *intrinseche*, su cosa sia fatto di essere in quegli stati. La distinzione è quella tra struttura e qualità, tra forma e contenuto fenomenico.

Questa osservazione è connessa alla critica che Bertrand Russell e Frank Ramsey avanzarono contro il riduzionismo strutturalista in filosofia della fisica: la fisica descrive le proprietà relazionali della materia — come le particelle si relazionano l'una all'altra, come interagiscono con i campi — ma non ci dice nulla delle proprietà intrinseche che le particelle possiedono al di là di queste relazioni. Il problema difficile della coscienza è, da questa prospettiva, un caso speciale del problema più generale dell'incompletezza strutturale della descrizione fisica: la fisica descrive le relazioni, non le qualità.

Nel caso del FEP, si può rendere questa incompletezza più precisa. L'energia libera F[q, o] è definita in termini di distribuzioni di probabilità — oggetti matematici astratti che possono essere realizzati su qualsiasi substrato fisico. Le equazioni del predictive coding descrivono la struttura delle interazioni tra unità di stato e unità di errore, ma sono insensibili alla natura intrinseca di quelle unità: che siano neuroni in un cervello biologico, transistor in un computer o valvole in una macchina pneumatica non fa differenza per le equazioni. Questa insensibilità al substrato è precisamente la realizzabilità multipla del funzionalismo — e genera lo stesso problema del funzionalismo: non spiega perché la realizzazione biologica sia accompagnata da qualia mentre (si suppone) la realizzazione su silicio non lo sia.

III. Il panpsichismo come risposta al problema delle proprietà intrinseche

Il problema delle proprietà intrinseche ha portato alcuni filosofi — in modo crescente negli ultimi anni — verso una forma di panpsichismo. La posizione, difesa con rigore da Philip Goff e con sfumature diverse da Galen Strawson e Thomas Nagel, sostiene che la coscienza o le proprietà proto-coscienti siano una caratteristica fondamentale e ubiqua della realtà fisica, non un epifenomeno emergente da substrati privi di esperienza.

L'argomento più sofisticato di Goff procede nel modo seguente. La fisica descrive le proprietà dei fondamentali fisici in termini puramente strutturali — equazioni che specificano le relazioni tra quantità, ma non il contenuto intrinseco di quelle quantità. Russell aveva notato che questa struttura relazionale deve essere sopportata da qualcosa — da proprietà intrinseche che istanziano le relazioni. La domanda è: che tipo di proprietà intrinseche sono quelle dei fondamentali fisici? Goff suggerisce che la risposta più parsioniosa è che queste proprietà intrinseche siano proto-coscienti o coscienti in una forma rudimentale — e che la coscienza umana complessa sia l'integrazione di queste proprietà elementari in sistemi sufficientemente organizzati.

Il panpsichismo non è una teoria del FEP, ma il FEP può essere messo in dialogo produttivo con esso. In una lettura panpsichista del FEP, la minimizzazione dell'energia libera sarebbe il meccanismo attraverso cui le proprietà proto-coscienti dei fondamentali fisici si organizzano in forme di esperienza via via più integrate e ricche. Il confine di Markov definirebbe le condizioni strutturali per una tale integrazione: soltanto i sistemi con un confine di Markov stabile — soltanto i sistemi che mantengono la propria identità nel tempo — possono integrare le proprietà proto-coscienti elementari in un'esperienza unitaria.

Questa lettura è speculativa ma non incoerente. Il suo punto debole principale è il problema della combinazione (*combination problem*): come si combinano le proto-esperienze elementari delle singole particelle per dare luogo all'esperienza unificata di un soggetto? Il FEP potrebbe in linea di principio suggerire una risposta: l'unificazione avviene attraverso la struttura gerarchica del modello generativo, che integra informazioni da molti sorgenti in un'unica rappresentazione coerente. Ma questa è una risposta funzionale al problema di combinazione, non una risposta fenomenologica: spiega come le informazioni si integrano strutturalmente, non come le proto-esperienze si fondono in un'unica esperienza.

IV. L'illusionismo e la dissoluzione del problema

All'estremo opposto del panpsichismo si trova l'illusionismo, difeso con acutezza da Keith Frankish. L'illusionismo sostiene che il problema difficile è un artificio cognitivo — un'illusione prodotta da una forma di auto-malinteso sistematico. Non esistono qualia nel senso in cui il problema difficile li postula: esistono stati funzionali che si *rappresentano* come aventi proprietà intrinseche irriducibili, ma questa auto-rappresentazione è inaccurata. Ciò che sembra essere la qualità intrinseca del rosso di una mela è in realtà la rappresentazione cerebrale di una proprietà funzionale dello stato visivo — una rappresentazione che attribuisce a quello stato proprietà che esso non possiede.

L'illusionismo è compatibile con il FEP e, in un certo senso, ne è suggerito. Se il sé è un costrutto inferenziale — se il cervello genera un modello del sé come parte del suo modello generativo — allora il senso di avere esperienze qualitative potrebbe essere anch'esso un contenuto del modello generativo piuttosto che una proprietà reale degli stati neurali. Il cervello si rappresenta come avente qualia, ma questa auto-rappresentazione potrebbe essere una caratteristica del modello generativo che non corrisponde a nessuna proprietà fenomenica intrinseca degli stati neurali stessi.

L'illusionismo ha il pregio dell'economia ontologica: elimina il problema difficile dichiarandolo mal posto. Ma il prezzo è alto — forse troppo alto per essere pagato con serenità. Il critico più acuto dell'illusionismo, che si può chiamare il critico fenomenologico, obietta nel modo seguente: l'illusionismo afferma che sembra esserci un'esperienza qualitativa, ma in realtà non c'è. Ma il sembrare è esso stesso un fatto fenomenico: se qualcosa *sembra* avere proprietà qualitative, questo sembramento *è* un'esperienza qualitativa. L'illusionismo, per evitare il problema difficile, deve eliminare non soltanto i qualia ma anche l'apparenza dei qualia — e questo è precisamente ciò che sembra impossibile eliminare senza contraddirsi.

Nel linguaggio del FEP, questa obiezione si traduce così: anche se il modello generativo del cervello si rappresenta come avente qualia, questa auto-rappresentazione è essa stessa un evento neurale che ha una struttura inferenziale. Ma perché questa auto-rappresentazione sia *un'illusione* — perché sembri qualcosa che non è — è necessario che ci sia qualcosa che sembra. E questo sembramento non può essere eliminato per decreto.

V. L'enattivismo e le sue consonanze con il FEP

Una terza posizione nel panorama contemporaneo è l'enattivismo, sviluppato in diverse varianti da Francisco Varela, Evan Thompson e Alva Noë, e più recentemente da Thomas Metzinger nella sua neurofenomenologia. L'enattivismo sostiene che la coscienza non è una proprietà interna del cervello isolato ma emerge dalla dinamica di accoppiamento tra organismo e ambiente: il soggetto *enattua* il suo mondo attraverso le sue capacità sensorimotrici, e l'esperienza è inseparabile da questo processo di enazione.

Le consonanze tra l'enattivismo e il FEP sono profonde e non casuali. L'inferenza attiva è enattivista nel suo spirito: il sistema non è un ricevitore passivo che costruisce rappresentazioni del mondo esterno, ma un agente che *genera* il proprio flusso sensoriale attraverso le proprie azioni. La percezione non precede l'azione: le due si co-determinano in un circolo continuo. Il confine di Markov, come interfaccia sensorimotrice tra sistema e ambiente, è la formalizzazione matematica di ciò che i teorici dell'enazione descrivono come il confine di accoppiamento strutturale tra organismo e nicchia.

Tuttavia, l'enattivismo nella sua versione più radicale va oltre il FEP in un punto cruciale: afferma che l'esperienza soggettiva è irriducibile alla struttura computazionale interna del sistema, perché emerge dalla relazione organismo-ambiente come proprietà del sistema allargato. Questa tesi richiede che il confine di Markov non sia il luogo in cui finisce l'esperienza soggettiva ma il luogo in cui essa prende vita — che i qualia siano proprietà del processo di accoppiamento e non dei soli stati interni.

Il FEP può accogliere questa idea parzialmente: la nozione di inferenza attiva è compatibile con l'idea che l'esperienza emerga dal circolo sensorimorale piuttosto che dagli stati interni isolati. Ma il FEP non può derivare da solo questa conclusione: può al massimo indicare che le condizioni strutturali per l'esperienza includono necessariamente l'accoppiamento con l'ambiente, senza spiegare perché questo accoppiamento sia esperito.

VI. L'integrazione dell'informazione e il dialogo con il FEP

Prima di trarre le conclusioni, è opportuno considerare brevemente il rapporto del FEP con la teoria dell'integrazione dell'informazione (Integrated Information Theory, IIT) di Giulio Tononi — l'altra grande candidata a teoria matematica della coscienza.

L'IIT propone che la coscienza sia identica all'integrazione dell'informazione nel sistema, misurata dalla quantità φ (phi): la quantità di informazione generata dal sistema come un tutto al di sopra e al di là di quella generata dalle sue parti. Formalmente, φ è definita come:

$$\varphi = \min_{P} KL[p(X^t \mid X^{t-1}) \;//\; \prod_i p(X_i^t \mid X_i^{t-1})]$$

dove la minimizzazione è sul taglio minimo di informazione P del sistema: la partizione del sistema in due parti che minimizza la quantità di informazione integrata. Un sistema con φ alto è altamente cosciente; un sistema con φ basso o nullo non ha esperienza.

Il FEP e l'IIT condividono l'ambizione di quantificare la struttura computazionale della coscienza con strumenti matematici precisi, ma hanno architetture molto diverse. L'IIT è essenzialmente bottom-up: parte dalla misura φ e deriva da essa le proprietà dell'esperienza (unità, struttura, specificità). Il FEP è essenzialmente top-down: parte dal principio di minimizzazione dell'energia libera e deriva da esso le proprietà strutturali della cognizione.

Un punto di contatto rilevante è il seguente. Si può mostrare che un sistema con confine di Markov stabile e modello generativo gerarchico tende a sviluppare stati interni con alta integrazione dell'informazione: la struttura gerarchica che permette la minimizzazione efficiente dell'energia libera è anche la struttura che massimizza l'integrazione dell'informazione attraverso i livelli. Questo non è una coincidenza: entrambe le proprietà — efficienza computazionale e integrazione — sono favorite da architetture modulari-gerarchiche.

Tuttavia, anche se φ e F fossero matematicamente correlate — anche se i sistemi con alto φ tendessero a essere buoni minimizzatori di energia libera e viceversa — questo non dissolverebbe il problema difficile per nessuna delle due teorie: resterebbe la domanda di perché l'alta integrazione o la bassa energia libera siano accompagnate da esperienza. La misura φ e la misura F sono entrambe misure di struttura funzionale, non di qualità fenomenica.

VII. Il contributo originale: la topologia dell'explanatory gap

Si vuole ora avanzare la tesi originale di questo saggio in forma più precisa. Si propone che il FEP, congiunto alla teoria del confine di Markov, contribuisca al problema difficile non risolvendo ma *topologizzando* l'explanatory gap: fornendo una descrizione formale della struttura geometrica dello spazio in cui il problema si colloca, individuando con precisione le dimensioni lungo le quali la spiegazione funzionale è completa e le dimensioni lungo le quali essa è intrinsecamente incompleta.

Più specificamente, si propone la seguente distinzione. Il FEP descrive completamente ciò che si può chiamare la *topologia relazionale* di un sistema cognitivo: la struttura delle relazioni causali tra stati interni, stati sensoriali, stati di azione e stati esterni; la dinamica dell'aggiornamento delle credenze; la gerarchia dei modelli generativi; la struttura temporale delle politiche; il confine che separa il sistema dall'ambiente. Questa topologia relazionale è completamente determinata dalla struttura del processo di minimizzazione dell'energia libera.

Ma il problema difficile riguarda ciò che si può chiamare la *metrica intrinseca* degli stati interni: la proprietà in virtù della quale un certo stato interno è il *tipo di esperienza* che è, con le sue specifiche qualità fenomeniche — la rossità del rosso, la dolcezza del dolce, il dolore del dolore. Questa metrica intrinseca non è determinata dalla topologia relazionale: sistemi con la stessa topologia relazionale potrebbero avere metriche intrinseche diverse (o nessuna metrica intrinseca), e sistemi con metriche intrinseche diverse potrebbero avere la stessa topologia relazionale.

L'analogia matematica che si vuole evocare è quella tra topologia e metrica in geometria differenziale. Due spazi possono avere la stessa topologia (essere omeomorfi) pur avendo metriche diverse: uno sfera e un ellissoide hanno la stessa topologia ma metriche diverse. La topologia determina la struttura globale, la metrica determina le proprietà locali e quantitative. Allo stesso modo, la topologia relazionale della cognizione (descritta dal FEP) e la metrica intrinseca dell'esperienza (i qualia) sono aspetti ortogonali della mente che non si determinano vicendevolmente.

Questa distinzione topologia/metrica permette di formulare la limitazione del FEP in modo preciso senza cadere né nel dualismo delle sostanze né nell'illusionismo: il FEP descrive la struttura topologica della mente in modo completo e rigoroso, ma è intrinsecamente silenzioso sulla struttura metrica — e il problema difficile è precisamente il problema della metrica intrinseca dell'esperienza.

VIII. Cosa dovrebbe aggiungere una teoria completa

Se questa diagnosi è corretta, quali sono le condizioni che una teoria completa della mente dovrebbe soddisfare, al di là di ciò che il FEP già fornisce?

La prima condizione è quella che si può chiamare il *principio di determinazione intrinseca*: la teoria deve specificare non soltanto la struttura relazionale degli stati interni di un sistema cognitivo ma anche le proprietà intrinseche che queste relazioni istanziano. Questo richiede un'estensione del formalismo che includa, in qualche modo, informazioni sul substrato fisico — non in modo da rinunciare alla realizzabilità multipla, ma in modo da riconoscere che substrati fisici diversi possono realizzare la stessa topologia relazionale con metriche intrinseche diverse.

La seconda condizione è quella del *principio di continuità fenomenologica*: la teoria deve essere in grado di derivare la struttura dell'esperienza soggettiva — la sua temporalità, la sua prospetticità, la sua unità — dalle sue condizioni di possibilità formali, non soltanto come proprietà emersa a posteriori ma come conseguenza necessaria della struttura del sistema. Il FEP si avvicina a questa condizione con la nozione di sé come costrutto inferenziale e con la struttura temporale delle politiche, ma non la soddisfa pienamente.

La terza condizione è quella del *principio di invarianza fenomenica*: la teoria deve essere in grado di spiegare perché sistemi con diversi substrati fisici ma identica topologia relazionale abbiano (o non abbiano) esperienze qualitativamente indistinguibili. Questo è il nucleo del problema della realizzabilità multipla applicato alla fenomenologia: se la coscienza è realizzabile su substrati multipli, cosa rende invariante la qualità fenomenica attraverso le realizzazioni?

Nessuna delle teorie attuali — né il FEP, né l'IIT, né l'enattivismo, né il panpsichismo — soddisfa tutte e tre queste condizioni. Ma il FEP, per la sua precisione matematica e la sua generalità formale, costituisce la base più solida su cui costruire la struttura che potrebbe, eventualmente, soddisfarle. È in questo senso, e non nel senso di aver risolto il problema

difficile, che il FEP rappresenta il progresso più importante nella comprensione della mente degli ultimi decenni.

Bilancio Critico e Indicazioni per una Teoria Futura

I. Ciò che il FEP ha conquistato

Qualsiasi valutazione critica onesta deve cominciare con un riconoscimento generoso: il Principio di Energia Libera ha trasformato irreversibilmente il paesaggio delle neuroscienze computazionali e della filosofia della mente. Non si tratta di una trasformazione meramente tecnica — dell'aggiunta di uno strumento matematico a una cassetta degli attrezzi già fornita — bensì di un riorientamento concettuale che ha cambiato il tipo di domande che si considera legittimo porre.

Prima del FEP, la questione dominante era: quali algoritmi implementa il cervello? Dopo il FEP, la domanda diventa: quale principio normativo governa il cervello come sistema adattativo? Il salto non è superficiale. Nel primo regime, ogni fenomeno cognitivo esige una spiegazione separata: la percezione visiva ha la sua architettura, l'attenzione ha la sua, la pianificazione motoria ha la sua. Nel secondo regime, tutti questi fenomeni sono manifestazioni di un unico principio applicato a strutture di modelli generativi diverse. L'unificazione che ne risulta non è soltanto esteticamente soddisfacente: genera predizioni comuni che si possono verificare trasversalmente, crea ponti tra campi disciplinari che prima si parlavano poco e permette di trattare con un unico formalismo fenomeni che sembravano irrelati — come la connessione tra dolore cronico e precisione interoceptiva aberrante, o quella tra creatività artistica e bassa precisione dei priori.

Il secondo contributo durevole del FEP è la formalizzazione del sé come costrutto inferenziale. Prima di Friston, l'idea che il sé fosse una costruzione — una narrativa, un'ipotesi, un'astrazione — era patrimonio della fenomenologia, della psicoanalisi e di certe correnti del buddismo, ma non aveva una formulazione matematica precisa. Il FEP fornisce questa formulazione: il sé è la componente del modello generativo che rappresenta il sistema come causa delle proprie azioni sensoriali, mantenuta dalla minimizzazione dell'energia libera rispetto a un confine di Markov stabile. Questa formulazione è operazionalizzabile, genera predizioni su come il senso del sé dovrebbe alterarsi in condizioni di alterazione del confine di Markov (stati dissociativi, esperienze fuori dal corpo, depersonalizzazione) e permette un dialogo produttivo tra neuroscienze e tradizioni contemplative orientali.

Il terzo contributo è la naturalizzazione dell'intenzionalità — della proprietà degli stati mentali di essere *di* qualcosa, di avere un contenuto che riguarda il mondo. Nel FEP, l'intenzionalità è la relazione tra il modello generativo e i suoi oggetti: gli stati nascosti ψ che il modello tenta di inferire. Questa relazione non è misteriosa: è la relazione matematica tra le variabili di un modello probabilistico e le distribuzioni che esse approssimano.

L'intenzionalità non richiede un'anima cartesiana che conferisce significato agli stati neurali: è una proprietà strutturale del processo di inferenza, determinata dalla forma del modello generativo e dalla sua relazione con i segnali sensoriali. Questo non dissolve il problema filosofico dell'intenzionalità nella sua interezza — la questione del contenuto fenomenico dell'intenzionalità resta — ma ne risolve la componente funzionale in modo convincente.

II. I limiti strutturali: una tassonomia

Accanto a questi contributi, il FEP presenta limitazioni che vanno distinte con cura perché hanno nature diverse e richiedono risposte diverse. Si propone qui una tassonomia in tre livelli.

Il primo livello comprende le limitazioni tecniche dell'approssimazione: i casi in cui il formalismo attuale non è abbastanza potente da catturare la complessità dei fenomeni che pretende di spiegare. L'approssimazione gaussiana che si è discussa nella Parte Seconda è il caso principale. Ma appartengono a questo livello anche la difficoltà di specificare a priori la struttura del modello generativo per sistemi biologici reali, la questione aperta dell'apprendimento dei parametri del modello nel corso dello sviluppo ontogenetico, e la tensione tra la natura continua del formalismo variazionale e la natura discreta di molte strutture cognitive (categorie, concetti, grammatica). Queste limitazioni sono in linea di principio sormontabili con sviluppi tecnici del formalismo — e di fatto molti ricercatori vi stanno lavorando, con risultati promettenti nel campo dei modelli generativi profondi e dell'apprendimento per rinforzo bayesiano.

Il secondo livello comprende le limitazioni epistemologiche: i casi in cui il FEP avanza pretese esplicative che le sue risorse concettuali non possono sostenere. La pretesa più prominente di questo tipo è l'affermazione che qualsiasi sistema con un confine di Markov possa essere descritto come agente che compie inferenza, e che questa descrizione sia più che una comoda finzione — che abbia valenza ontologica. Come si è discusso nella Parte Terza, questa pretesa è eccessiva: l'equivalenza formale tra la dinamica di un sistema e le equazioni dell'inferenza bayesiana non implica che il sistema *davvero* inferisca nel senso normativo e rappresentazionale del termine. Friston è a volte ambiguo su questo punto, oscillando tra una lettura strumentalista (il formalismo è un utile strumento descrittivo) e una lettura realista (il formalismo descrive ciò che il sistema fa realmente). La distinzione è filosoficamente cruciale e merita maggiore chiarezza di quella che si trova nella letteratura primaria.

Il terzo livello è quello che si è chiamato la limitazione intrinseca: il caso in cui il FEP non soltanto non spiega qualcosa ma non può spiegarlo in linea di principio, perché la domanda riguarda dimensioni della realtà che il formalismo non cattura strutturalmente. Il problema difficile della coscienza appartiene a questo livello, come si è argomentato nella Parte Quarta. Non si tratta di una lacuna tecnica da colmare: si tratta di un limite costitutivo, nella stessa misura in cui la meccanica classica non può spiegare i fenomeni quantistici non perché le sue equazioni siano abbozzate ma perché le sue variabili fondamentali non includono il grado di libertà rilevante.

Questa tassonomia è importante perché impedisce un errore simmetrico e opposto a quello dell'entusiasmo eccessivo: l'errore del rigetto globale. Chi critica il FEP adducendo che non risolve il problema difficile spesso commette il rovescio dell'errore dell'entusiasta: attribuisce al FEP pretese che esso non avanza (o non dovrebbe avanzare) nelle sue versioni più sobrie, e poi lo critica per non soddisfarle. La critica corretta è più chirurgica: riconosce cosa il FEP spiega genuinamente, isola le pretese che eccedono le sue risorse concettuali, e identifica le domande per cui un formalismo diverso è necessario.

III. Il problema della circolarità e la questione del bootstrapping

Un'obiezione al FEP che si incontra raramente nella letteratura, ma che merita attenzione, riguarda una forma sottile di circolarità nel fondamento del principio stesso. Il FEP afferma che qualsiasi sistema che sopravviva nel tempo deve minimizzare l'energia libera — che la sopravvivenza *è* minimizzazione dell'energia libera. Ma come viene giustificata questa identità?

L'argomento di Friston procede grosso modo così. I sistemi biologici abitano nicchie ecologiche specifiche: si trovano in un insieme limitato di stati compatibili con la loro sopravvivenza. Questo significa che la distribuzione di probabilità dei loro stati sensoriali è concentrata in una regione ristretta dello spazio degli stati — hanno bassa entropia marginale. Poiché l'energia libera F è un limite superiore della sorpresa -ln p(o), e la sorpresa media di un sistema con bassa entropia sensoriale è bassa, un sistema che sopravvive deve avere energia libera bassa in media. Quindi la sopravvivenza implica la minimizzazione dell'energia libera.

Questo argomento è valido, ma è più debole di quanto sembri. In primo luogo, mostra che i sistemi sopravvissuti hanno avuto bassa energia libera in media nel passato, non che abbiano un meccanismo che minimizza attivamente l'energia libera nel futuro. La distinzione è tra una proprietà statistica dei sistemi sopravvissuti (selezione naturale per la bassa sorpresa) e un meccanismo attivo di minimizzazione (il cervello come ottimizzatore variazionale). Il primo è compatibile con molti meccanismi diversi; il secondo è la tesi specifica del FEP.

In secondo luogo, l'argomento presuppone che il modello generativo del sistema sia accurato — che $p(o \mid \psi)$ nel modello rifletta davvero le statistiche del mondo. Ma i sistemi biologici hanno modelli generativi appresi nel corso dell'evoluzione e dello sviluppo, e questi modelli possono essere sistematicamente distorti. Un sistema con un modello generativo parzialmente scorretto può sopravvivere minimizzando l'energia libera sotto quel modello, non perché il suo modello sia fedele al mondo ma perché il mondo e il modello sono co-evoluti fino a corrispondersi sufficientemente. Questo introduce una forma di circolarità: il modello generativo è definito rispetto al mondo, ma il mondo rilevante per il sistema è co-costruito dal comportamento del sistema, che è guidato dal modello generativo.

Questa circolarità non è viciosa — è precisamente la circolarità dell'accoppiamento strutturale che l'enattivismo pone al centro della cognizione. Ma significa che il FEP non è un principio fondato in modo puramente a priori: dipende da assunzioni empiriche sulla

struttura del mondo e sulla storia evolutiva del sistema. Non è un assioma della vita ma una proprietà dei sistemi che hanno avuto successo in particolari nicchie ecologiche.

IV. Geometria dell'esperienza: abbozzi di una teoria futura

Si tenta ora di abbozzare, con la prudenza dovuta, le linee di una teoria futura che potrebbe sopperire ai limiti che si sono identificati, senza pretendere di aver risolto problemi che rimangono aperti.

La proposta si articola attorno alla nozione di *geometria fenomenica* — un formalismo che assegni agli stati interni di un sistema cognitivo non soltanto una struttura topologica (descritta dal FEP) ma anche una struttura metrica intrinseca che corrisponda alle proprietà qualitative dell'esperienza.

Il punto di partenza è un'osservazione sulla struttura dello spazio fenomenico dei colori. I colori non sono semplicemente discriminabili o non discriminabili: si somigliano e differiscono in modi che hanno una struttura metrica precisa. Il rosso somiglia all'arancione più di quanto somiglia al blu; il rosa somiglia al bianco più di quanto somiglia al rosso saturo. Questa struttura è catturata dallo spazio dei colori di Munsell o dallo spazio CIELAB — spazi metrici a tre dimensioni in cui le distanze corrispondono a giudizi di somiglianza percettiva. Questa struttura metrica non è determinata dalla struttura topologica delle connessioni neurali: è una proprietà aggiuntiva che deve essere spiegata da una teoria completa.

Si propone che le proprietà metriche dello spazio fenomenico siano in corrispondenza con le proprietà metriche dello spazio dei modelli generativi — non la topologia di quel spazio, che è la struttura delle relazioni tra stati del modello, bensì la curvatura locale di quel spazio, misurata dalla metrica di Fisher-Rao. La metrica di Fisher-Rao $g_{ij}(\theta)$ su una famiglia di distribuzioni $p(x\,;\theta)$ parametrizzata da θ è definita come:

$$g_{ij}(\theta) = \mathbb{E}_{p(x;\theta)}\left[\left(\partial \ln p(x;\theta)/\partial\theta_i\right) \cdot \left(\partial \ln p(x;\theta)/\partial\theta_j\right) \right]$$

ossia la matrice di informazione di Fisher. Questa metrica misura quanto velocemente cambia la distribuzione $p(x\,;\theta)$ al variare dei parametri θ: due punti nello spazio dei parametri sono vicini secondo la metrica di Fisher se e solo se le distribuzioni corrispondenti sono difficilmente distinguibili statisticamente. La distanza di Fisher-Rao è la distanza geodetica nello spazio statistico delle distribuzioni — la minima quantità di informazione necessaria per passare da una distribuzione all'altra.

La congettura che si avanza è la seguente: le distanze fenomeniche nello spazio qualitativo dell'esperienza (quanto due qualia differiscono qualitativamente) sono proporzionali alle distanze di Fisher-Rao nello spazio dei parametri del modello generativo. Due stati fenomenici si somigliano qualitativamene nella misura in cui le distribuzioni del modello generativo corrispondenti sono vicine secondo la metrica di Fisher.

Formalmente, se θ_1 e θ_2 sono due punti nello spazio dei parametri del modello generativo, e Q(θ_1) e Q(θ_2) sono i due stati fenomenici corrispondenti, si postula:

$$d_{\text{fenomenica}}(Q(\theta_1), Q(\theta_2)) \propto d_{\text{Fisher}}(\theta_1, \theta_2)$$

dove d_{Fisher} è la distanza geodetica secondo la metrica di Fisher-Rao:

$$d_{\text{Fisher}}(\theta_1, \theta_2) = \min_{\gamma} \int_0^1 \sqrt{\dot{\gamma}(t)^T \, g(\gamma(t)) \, \dot{\gamma}(t)} \, dt$$

con $\gamma(t)$ una curva che connette θ_1 a θ_2 nello spazio dei parametri.

Questa congettura ha diverse virtù. In primo luogo, è misurabile: la metrica di Fisher-Rao è calcolabile per modelli generativi espliciti, e le distanze fenomeniche possono essere misurate psicofisicamente tramite giudizi di somiglianza. Se la congettura è vera, ci si aspetta una correlazione tra le due misure. In secondo luogo, è compatibile con il FEP: la metrica di Fisher è la metrica naturale dello spazio in cui opera l'inferenza variazionale — i passi di aggiornamento delle credenze che minimizzano l'energia libera percorrono geodetiche in questo spazio quando il sistema è in regola naturale (natural gradient). In terzo luogo, estende il FEP senza contraddirlo: aggiunge una struttura metrica allo spazio dei parametri senza alterare la struttura topologica.

L'obiezione ovvia è che questa congettura non risolve il problema difficile: anche se le distanze fenomeniche corrispondessero alle distanze di Fisher, resterebbe la domanda di perché a questa corrispondenza sia associata l'esperienza soggettiva piuttosto che nulla. La risposta, nella prospettiva di questo saggio, è che la congettura non mira a risolvere il problema difficile ma a formulare una condizione di adeguatezza per qualsiasi sua soluzione: qualsiasi teoria che soddisfi il principio di determinazione intrinseca introdotto nella Parte Quarta deve essere in grado di derivare la struttura metrica dello spazio fenomenico, e la congettura specifica quale struttura metrica è rilevante — quella di Fisher-Rao nello spazio dei modelli generativi.

V. Il naturalismo non riduzionista come esito

Il percorso di questo saggio conduce a una posizione che si può chiamare naturalismo non riduzionista rispetto alla coscienza. Questa posizione sostiene le seguenti tesi, che si articolano in ordine di forza decrescente:

La tesi forte: la struttura funzionale della cognizione — percezione, azione, apprendimento, attenzione, sé — è completamente spiegata dal FEP e dalle sue estensioni. Non è necessario postulare alcun principio non-fisico per spiegare come un sistema cognitivo funziona.

La tesi moderata: la struttura dell'esperienza soggettiva — la sua organizzazione qualitativa, le relazioni di somiglianza tra qualia, la sua struttura temporale — è connessa in modo sistematico e probabilmente necessario alla struttura del modello generativo del sistema, e

questa connessione può essere progressivamente chiarita da teorie come quella abbozzata nella sezione precedente.

La tesi debole: l'esistenza dell'esperienza soggettiva — il fatto che ci sia qualcosa che è come essere un certo sistema — non è derivabile a priori dalla struttura fisica o computazionale del sistema, e rimane un problema aperto che il naturalismo deve affrontare senza poterlo risolvere per ora.

Questa triplice struttura è compatibile con il realismo sulle proprietà fenomeniche senza richiedere il dualismo delle sostanze. Non è necessario postulare una sostanza mentale separata da quella fisica: è sufficiente riconoscere che le proprietà fisiche dei sistemi neurali includono proprietà fenomeniche intrinseche che non si riducono alle loro proprietà relazionali-funzionali. Questa è la posizione che in filosofia analitica si chiama monismo delle proprietà duali o, nella versione più vicina al pensiero russelliano, monismo neutro.

Il FEP è compatibile con questa posizione e ne fornisce l'impalcatura funzionale: descrive completamente il lato funzionale della mente, che è il lato che la scienza può investigare con i propri metodi. Il lato fenomenico rimane, per ora, il territorio inesplorato sulla mappa — non perché sia soprannaturale o inconoscibile per principio, ma perché i concetti e i metodi necessari per esplorarlo non sono ancora stati sviluppati con sufficiente rigore.

VI. La domanda che resta

Ogni saggio filosofico di un certo respiro dovrebbe concludersi non con una risposta ma con la domanda più acuta che è riuscito a formulare nel corso del cammino. La domanda che questo saggio ha cercato di affinare è la seguente.

Il FEP mostra che qualsiasi sistema che mantenga la propria identità nel tempo è, formalmente, un sistema che minimizza una misura di sorpresa rispetto a un modello del mondo — un sistema che *preferisce* certi stati a certi altri, in un senso precisamente definibile in termini probabilistici. Questa preferenza non è una metafora: è una proprietà fisica, derivabile dalla struttura del processo stocastico.

Ora, c'è una differenza tra preferire nel senso fisico del FEP e preferire nel senso fenomenico dell'esperienza soggettiva: la prima è una proprietà strutturale di certi sistemi dinamici, la seconda implica che ci sia qualcosa che è come avere quella preferenza — un sentire, un essere attratti o respinti. Il problema difficile si riduce infine a questa domanda: cosa trasforma la preferenza strutturale in preferenza vissuta?

Il FEP non risponde a questa domanda. Ma ha il merito, non piccolo, di averla formulata con una chiarezza che prima non aveva.

Psicopatologia Computazionale e Filosofia dell'Azione

I. La mente che si rompe come specchio della mente che funziona

Esiste nella storia della medicina e della filosofia una lunga tradizione che utilizza la patologia come strumento euristico per comprendere la normalità. La neuropsicologia classica, da Broca e Wernicke in poi, ha costruito la sua mappa delle funzioni cognitive studiando i deficit prodotti da lesioni focali; la psicopatologia fenomenologica, da Jaspers a Minkowski a Binswanger, ha illuminato la struttura dell'esperienza normale studiando le sue forme di disintegrazione nella malattia mentale. Il FEP si inserisce in questa tradizione con una proposta teorica di insolita precisione: i disturbi mentali possono essere compresi come alterazioni specifiche e localizzabili nel processo di minimizzazione dell'energia libera — come distorsioni del modello generativo, aberrazioni nella stima della precisione, o disfunzioni nella struttura del confine di Markov.

Questa non è una metafora: è una proposta empiricamente verificabile. Se il FEP è la teoria corretta del funzionamento cerebrale, allora la psicopatologia computazionale — la disciplina che studia i disturbi mentali come disturbi del processo inferenziale — dovrebbe essere in grado di generare predizioni specifiche su quali parametri del modello generativo sono alterati in ciascun disturbo, predizioni verificabili con misure comportamentali, neurofisiologiche e di neuroimmagine.

II. La psicosi come aberrazione della precisione

La proposta più sviluppata e empiricamente articolata nella letteratura del FEP riguarda la psicosi, e in particolare la schizofrenia. L'ipotesi, sviluppata da Philip Corlett, Karl Friston, Paul Fletcher e colleghi nel corso degli anni 2010, sostiene che i sintomi psicotici emergono da una disregolazione della precisione — una perturbazione nella ponderazione relativa dei segnali di errore di predizione rispetto alle predizioni del modello.

Per articolare questa proposta con precisione, si richiami la struttura del predictive coding descritta nella Parte Seconda. Il segnale di errore di predizione ponderato dalla precisione al livello l è:

$$\xi^{(l)} = \Pi^{(l)} \cdot \varepsilon^{(l)} = \Pi^{(l)} \cdot [\mu^{(l)} - f^{(l)}(\mu^{(l+1)})]$$

dove $\Pi^{(l)}$ è la matrice di precisione (la stima della affidabilità del segnale sensoriale o del modello al livello l). In un sistema che funziona normalmente, la precisione è calibrata in modo adattivo: i segnali sensoriali affidabili (basso rumore) ricevono alta precisione, le predizioni robuste (bassa incertezza) ricevono alta precisione. L'aggiornamento delle credenze è proporzionale alla discrepanza pesata dalla precisione.

L'ipotesi psicosi-come-aberrazione-della-precisione afferma che nella schizofrenia la stima della precisione è sistematicamente distorta in direzione dell'iperattivazione: gli errori di predizione ricevono peso eccessivo, indipendentemente dalla loro reale affidabilità. Questo corrisponde, nella fenomenologia clinica, all'esperienza tipicamente prodromica di *apofenia* — la percezione di connessioni significative tra eventi casuali, il senso che eventi ordinari abbiano un significato speciale e destinato precisamente al paziente. Se ogni piccolo errore di predizione viene amplificato dalla precisione aberrante, il sistema interpreta ogni discrepanza minima come evidenza di una causa nascosta significativa, producendo la cascata inferenziale che porta alla formazione del delirio.

In termini matematici, se la precisione al livello sensoriale è $\Pi^{(o)} >> \Pi^{(o)}_{normale}$, l'aggiornamento del modello di livello superiore in risposta a un errore sensoriale ε è amplificato:

$$\Delta\mu^{(1)} \propto (\partial f^{(1)}/\partial\mu^{(1)})^{T} \cdot \Pi^{(o)} \cdot \varepsilon^{(o)}$$

e se questo aggiornamento spinge $\mu^{(1)}$ verso valori anomali, la cascata di aggiornamento si propaga verso l'alto nella gerarchia, distorcendo progressivamente le credenze a tutti i livelli. Il delirio, in questa lettura, è la migliore ipotesi che un sistema con precisione aberrante riesce a costruire per spiegare il proprio flusso sensoriale amplificato — non un difetto della logica inferenziale, bensì una inferenza razionalmente coerente a partire da premesse distorte.

Il substrato neurologico proposto per questa aberrazione della precisione è la disfunzione dopaminergica. La dopamina, in questa ipotesi, segnala la precisione attesa degli errori di predizione: livelli elevati di dopamina corrispondono a elevata precisione attesa, e quindi ad una maggiore suscettibilità all'aggiornamento delle credenze in risposta a segnali sensoriali. L'iperdopaminergia mesolimbica che si osserva nella psicosi acuta — e che è il meccanismo d'azione dei farmaci antipsicotici che bloccano i recettori D2 — è precisamente il correlato proposto dell'iperprecisione aberrante. Questa lettura è compatibile con la cosiddetta ipotesi del salience della dopamina di Shitij Kapur, ma la embedded in un formalismo molto più preciso che permette predizioni quantitative.

I sintomi negativi della schizofrenia — apatia, anedonia, ritiro sociale, impoverimento del pensiero — possono essere interpretati, in questa stessa cornice, come la risposta adattiva del sistema a un'esperienza di errori di predizione amplificati e incontrollabili: un abbassamento difensivo della precisione sensoriale, una riduzione della suscettibilità agli aggiornamenti, che riduce l'angoscia dell'apofenia ma al costo di disconnettere progressivamente il sistema dal flusso informativo del mondo.

III. La depressione come priore eccessivamente preciso

La depressione maggiore può essere compresa, nel formalismo del FEP, attraverso un meccanismo duale rispetto a quello della psicosi: non un'iperprecisione dei segnali di errore ma un'iperprecisione dei priori — una rigidità patologica delle aspettative che impedisce l'aggiornamento flessibile delle credenze in risposta all'evidenza.

Formalmente, sia p^D(ψ) la distribuzione a priori di un soggetto depresso sugli stati del mondo. L'ipotesi è che questa distribuzione sia caratterizzata da una precisione elevata (bassa varianza) concentrata in stati negativi:

p^D(ψ) ≈ $\mathcal{N}$(μ_negativo, Σ_bassa)

dove μ_negativo è un valore medio che rappresenta stati del mondo negativi (pericolo, perdita, fallimento) e Σ_bassa è una covarianza piccola — il soggetto è *certo* che il mondo sia negativo. Con una distribuzione a priori così precisa e negativa, l'aggiornamento bayesiano in risposta a evidenza positiva è soppresso: la verosimiglianza di un evento positivo è interpretata come anomalia, come errore di percezione, piuttosto che come ragione per aggiornare le credenze generali. Il soggetto depresso non è irrazionale: compie inferenza bayesiana corretta sotto un priore distorto.

L'energia libera variazionale di questo sistema è:

F = KL[q(ψ) // p^D(ψ)] - $\mathbb{E}$_q[ln p(o | ψ)]

Il termine KL è piccolo perché q è già vicina al priore rigido p^D, ma il secondo termine — la precisione del modello sulle osservazioni — è anche esso basso, poiché il modello generativo distorto non riesce a spiegare adeguatamente le osservazioni positive. Il sistema è intrappolato in un minimo locale dell'energia libera da cui non può uscire perché ogni evidenza positiva viene interpretata come anomalia a bassa precisione piuttosto che come ragione per aggiornare le aspettative.

Questa lettura ha implicazioni per la psicoterapia cognitiva: le tecniche di ristrutturazione cognitiva possono essere interpretate come interventi che aumentano la precisione delle osservazioni positive (riducendo il bias di conferma negativo) o che abbassano la precisione del priore (ammorbidendo la certezza delle aspettative negative), entrambi effetti che riducono il divario tra osservazioni e modello e permettono al sistema di uscire dal minimo locale. Le terapie farmacologiche con antidepressivi serotoninergici possono essere interpretate, in questa cornice, come modulatori della precisione dei priori interoceptivi — altering the signal-to-noise ratio of emotional prediction errors.

L'aspetto più interessante di questa proposta è che suggerisce una misura neuroimmagistica della depressione: se la rigidità del priore è reale, si dovrebbe osservare una riduzione dell'ampiezza delle risposte agli errori di predizione positivi (nelle misurazioni EEG come la mismatch negativity e le P300) e un abbassamento della suscettibilità all'aggiornamento delle credenze in compiti decisionali sequenziali. Queste predizioni hanno trovato supporto empirico parziale, ma la verifica sistematica rimane un programma di ricerca aperto.

IV. L'autismo come riduzione dell'assunzione contestuale

Il disturbo dello spettro autistico offre un terzo caso di studio particolarmente illuminante, perché la proposta del FEP — sviluppata soprattutto da Pawan Sinha, Rebecca Lawson e Karsten Friston — affronta direttamente alcune caratteristiche fenomenologiche dell'autismo che le teorie precedenti faticavano a spiegare unitariamente: l'ipersensibilità sensoriale, la preferenza per la routine, la difficoltà nella generalizzazione, l'eccellenza in compiti di discriminazione percettiva locale, la difficoltà nella cognizione sociale.

La proposta centrale è che l'autismo sia caratterizzato da una riduzione della capacità di costruire priori contestuali forti — priori che modulano la precisione attesa dei segnali sensoriali in funzione del contesto. In un sistema tipicamente sviluppato, la corteccia di livello superiore genera priori contestuali che sopprimono la precisione degli errori di predizione a basso livello: si sa che ci si trova in una stanza tranquilla, quindi i rumori lievi vengono trattati come non significativi e soppressi. In un sistema con autismo, questa contestualizzazione è ridotta: la corteccia superiore fornisce priori contestuali meno potenti, e gli errori di predizione a basso livello ricevono un peso relativo maggiore.

Formalmente, sia $\Pi^{ctx}(o)$ la componente della matrice di precisione sensoriale dovuta alla contestualizzazione top-down. In un sistema tipico:

$$\Pi^(o)_totale = \Pi^(o)_bottom-up + \Pi^{ctx}(o)_top-down$$

dove il secondo termine fornisce una modulazione dipendente dal contesto. Se $\Pi^{ctx}(o)_top-down \approx 0$, la precisione sensoriale è determinata esclusivamente dalla struttura bottom-up del segnale, senza la regolazione contestuale. Il risultato è un sistema che tratta ogni segnale sensoriale con pari peso indipendentemente dal contesto — più accurato nella discriminazione locale (perché non sopprime i dettagli), ma meno capace di estrarre regolarità globali e contestuali.

Questa proposta unifica in modo elegante le caratteristiche fenomenologiche dell'autismo. L'ipersensibilità sensoriale emerge dall'assenza di soppressione contestuale dei segnali a basso livello. La preferenza per la routine emerge dal fatto che la routine riduce gli errori di predizione in assenza di priori contestuali forti: se non si ha la capacità di contestualizzare la novità, la novità genera errori di predizione amplificati e quindi angoscia. L'eccellenza nella discriminazione percettiva locale emerge dall'alta precisione dei segnali sensoriali non soppressi. La difficoltà nella cognizione sociale emerge dal fatto che l'interpretazione delle espressioni facciali, dei toni di voce e dei gesti richiede priori contestuali sofisticati e flessibili — esattamente il tipo di modellizzazione che il sistema autistico fatica a costruire.

Il valore di questa proposta non risiede soltanto nella sua potenza esplicativa unificante: risiede anche nelle implicazioni etiche che genera. Se l'autismo è una diversa configurazione del sistema di precisione piuttosto che un difetto, allora le caratteristiche cognitive dell'autismo hanno un valore intrinseco — in determinati ambienti e compiti, la configurazione autistica è superiore a quella tipica. Questo è compatibile con la filosofia della neurodiversità senza incorrere nell'errore opposto di negare che certi aspetti dell'autismo possano causare sofferenza reale.

V. La filosofia dell'azione rivisitata

Il FEP ha implicazioni altrettanto profonde per la filosofia dell'azione — quel settore della filosofia analitica che studia la struttura dei comportamenti intenzionali, la distinzione tra azioni e meri movimenti, e il problema della causalità mentale.

La questione fondamentale nella filosofia dell'azione è quella dell'agentività: cosa distingue un'azione da un'occorrenza causalmente prodotta? Davidson sosteneva che un evento è un'azione se è causato nel modo giusto da una credenza e un desiderio appropriati — la cosiddetta *causal theory of action*. Questo però genera il problema della devianza causale: ci sono casi in cui un evento è causato da credenze e desideri pertinenti ma non è un'azione (come nei casi di tremito nervoso che produce accidentalmente il risultato desiderato). La teoria causale deve specificare il tipo di connessione causale rilevante, e questa specificazione si rivela sorprendentemente difficile.

Il FEP offre una via diversa verso questo problema. Nell'inferenza attiva, la distinzione tra azione e occorrenza non è una distinzione tra tipi di eventi: è una distinzione strutturale nell'architettura del sistema causale che produce l'evento. Un'azione è un cambiamento nel mondo prodotto da stati interni del sistema attraverso gli stati di azione — ossia attraverso il confine di Markov — in modo tale da ridurre l'errore di predizione propriocettivo. Un'occorrenza è un cambiamento nel mondo non mediato da questa struttura.

Formalmente, sia a(t) il vettore di stati di azione al tempo t. Un evento motorio è un'azione nel senso del FEP se:

$$a(t) = \text{argmin}_a F[q, o(a)]$$

ossia se è il risultato della minimizzazione dell'energia libera rispetto agli stati di azione. Un evento motorio che non soddisfa questa condizione — un riflesso non-controllato, un tremito, un tic — non è un'azione nel senso del FEP.

Ma questa caratterizzazione è troppo restrittiva? Sembra escludere le azioni impulsive, le abitudini automatiche, i comportamenti semiconsapevoli che costituiscono la gran parte della vita motoria quotidiana. La risposta del FEP è che queste azioni *sono* minimizzazioni dell'energia libera, ma a livelli diversi della gerarchia: le azioni abituali sono mediate da priori di policy fortemente concentrati su politiche particolari, sviluppati attraverso l'apprendimento ripetuto fino a diventare quasi-automatici. Non richiedono un deliberazione esplicita perché la struttura del modello generativo le rende le politiche a più bassa energia libera attesa nella situazione corrente — le realizza come previsioni proprie-cettive che vengono soddisfatte quasi riflessivamente.

Questo ha una conseguenza rilevante per il dibattito sul libero arbitrio. La distinzione tra azione deliberata e azione automatica non è una distinzione tra processi consapevoli e processi causalmente determinati: nel FEP, entrambi sono causalmente determinati dal processo di minimizzazione. La differenza è che l'azione deliberata coinvolge i livelli superiori della gerarchia — i priori a lungo termine, i valori, le aspettative sui risultati — mentre l'azione automatica è dominata dai livelli inferiori, con priori di policy locali e temporalmente

prossimi. Il libero arbitrio, in questa lettura, non è l'assenza di determinazione causale: è il tipo di determinazione causale in cui i livelli superiori della gerarchia — quelli che in un'ottica personale si chiamano valori, preferenze, carattere — hanno peso decisivo nel selezionare le politiche.

VI. L'agency e il problema della causazione mentale

Un problema più profondo che la filosofia dell'azione eredita dal dualismo cartesiano è quello della causazione mentale: come possono gli stati mentali — credenze, desideri, intenzioni — causare eventi fisici nel corpo e nel mondo? Se gli stati mentali sono identici a stati fisici del cervello (come afferma il fisicalismo), questa domanda sembra dissolversi; ma la dissoluzione è più apparente che reale, perché il problema si sposta: come può uno stato fisico del cervello causare un altro evento fisico *in virtù delle sue proprietà mentali* — in virtù di ciò che esso rappresenta piuttosto che in virtù di come è fisicamente costituito?

Questo è il problema dell'esclusione causale di Jaegwon Kim, e il FEP lo affronta, anche se parzialmente, attraverso il concetto di causa formale. Nel formalismo variazionale, gli stati interni del sistema non causano le azioni direttamente: esse co-determinano le azioni come partecipanti in un processo di ottimizzazione globale. Le azioni non sono l'effetto di credenze-cause: sono la soluzione a un problema di ottimizzazione vincolata in cui le credenze entrano come parametri del funzionale obiettivo. In questa struttura, le proprietà rappresentazionali degli stati interni — il loro contenuto inferenziale — sono rilevanti per il tipo di azione prodotta non causalmente nel senso umeo ma formalmente: determinano la forma del funzionale che viene minimizzato.

Questa non è una risposta definitiva al problema della causazione mentale, ma è un riorientamento che apre prospettive interessanti. Suggerisce che il concetto di causazione efficiente, mutuato dalla meccanica classica, potrebbe non essere il concetto giusto per descrivere come la mente muove il corpo: potrebbe essere necessario recuperare, in un quadro rigorosamente naturalistico, qualcosa che assomiglia alla causazione formale aristotelica — la forma del processo come determinante del suo esito, piuttosto che la spinta di un evento su un altro evento.

VII. Coda: il meraviglioso problema

Questo saggio è cominciato con la constatazione che esiste un problema che si rifiuta di cedere. Si è percorsa la struttura matematica del FEP con sufficiente dettaglio da apprezzarne sia la potenza sia i limiti; si sono esaminate le principali posizioni filosofiche in dialogo con esso; si è avanzata la proposta originale della distinzione topologia/metrica e della geometria fenomenica basata sulla distanza di Fisher-Rao; si è esplorata la psicopatologia computazionale come territorio in cui il FEP trova alcune delle sue applicazioni più concrete e verificabili; e si è rivisitata la filosofia dell'azione alla luce dell'inferenza attiva.

Al termine di questo percorso, il problema difficile della coscienza resta irrisolto. Questa persistenza non è frustrante: è filosoficamente preziosissima. Un problema che si dissolve facilmente era un problema mal posto. Un problema che resiste alle teorie più sofisticate della mente che l'umanità abbia mai prodotto è un problema che tocca qualcosa di reale e profondo — qualcosa che appartiene alla struttura della realtà in un modo che le nostre categorie attuali non riescono ancora a catturare.

Il merito del FEP, in ultima analisi, è di aver trasformato questo problema da nebuloso a nitido. Prima del FEP, si poteva dibattere all'infinito su cosa esattamente mancasse alle spiegazioni funzionali della coscienza, perché le spiegazioni funzionali erano esse stesse imprecise. Dopo il FEP, il problema è localizzato con chirurgica precisione: ciò che manca è la metrica intrinseca, la specificazione delle proprietà fenomeniche che la topologia relazionale non determina. Sapere esattamente dove si trova un problema è il primo e forse il più importante passo verso la sua soluzione.

In questo senso, forse, il contributo più durevole del Principio di Energia Libera alla filosofia della mente non è la risposta che fornisce, bensì la domanda che rende finalmente possibile formulare con rigore: non «perché il cervello genera esperienza?» — domanda troppo vaga per essere fruttuosa — ma «qual è la struttura metrica degli stati interni di un sistema minimizzante l'energia libera, e come questa struttura si traduce nella geometria dello spazio fenomenico?». A questa domanda, la risposta non è ancora nota. Ma è una domanda con la forma giusta — una domanda che ammette, almeno in principio, una risposta empiricamente ancorata. E nella storia della scienza e della filosofia, formulare la domanda con la forma giusta è quasi sempre il preludio alla risposta.

Note Critiche, Evidenze Sperimentali e Confronti Teorici

I. Il predictive coding sotto il microscopio sperimentale

Una teoria tanto ambiziosa quanto il FEP non può permettersi di restare nell'empireo della pura speculazione matematica: deve confrontarsi con i dati, e il confronto deve essere abbastanza serrato da rischiare la falsificazione. La domanda è se le predizioni derivate dall'architettura del predictive coding gerarchico siano effettivamente verificate dalla neurofisiologia e dalla neuroimmagine, e in che misura l'evidenza supporti davvero il meccanismo specifico piuttosto che una classe più ampia di modelli compatibili.

Il paradigma sperimentale più direttamente rilevante è quello delle risposte alle violazioni delle aspettative — il cosiddetto paradigma oddball, in cui stimoli standard ripetuti vengono occasionalmente intervallati da stimoli devianti. Se il cervello genera predizioni e trasmette verso l'alto soltanto gli errori di predizione, uno stimolo deviante dovrebbe produrre una risposta neurale amplificata proporzionale all'entità della violazione delle aspettative. La risposta elettrofisiologica nota come Mismatch Negativity (MMN), una componente negativa del potenziale evento-correlato che compare circa 100-200 millisecondi dopo uno stimolo

deviante, è interpretata come il correlato neurale dell'errore di predizione uditivo. La struttura della MMN corrisponde qualitativamente alle predizioni del predictive coding: la sua ampiezza è proporzionale al grado di devianza, è presente anche in soggetti anestetizzati (suggerendo che non richieda attenzione consapevole), e la sua distribuzione scalpo è compatibile con l'attivazione delle aree uditive primarie e secondarie.

Più direttamente rilevante per l'architettura gerarchica è l'analisi della connettività effettiva tra aree corticali durante compiti percettivi, condotta con metodi come il Dynamic Causal Modelling (DCM) — uno strumento sviluppato da Friston per inferire la struttura delle connessioni causali tra regioni cerebrali da dati di risonanza magnetica funzionale. I risultati di numerosi studi DCM mostrano connessioni discendenti modulatorie compatibili con il trasporto di predizioni e connessioni ascendenti compatibili con il trasporto di errori di predizione, con la modulazione della connettività dipendente dall'incertezza stimata — come predice il meccanismo di regolazione della precisione.

È necessario tuttavia segnalare con franchezza i limiti metodologici. Il problema fondamentale è che il predictive coding non è l'unica architettura compatibile con la presenza di connessioni feedback modulatorie e segnali di errore nelle aree sensoriali primarie. Molte alternative — modelli di amplificazione del contrasto, normalizzazione divisiva, soppressione della ripetizione — producono predizioni simili nei paradigmi standard. Distinguere tra queste architetture richiede paradigmi appositamente progettati per discriminare le loro predizioni divergenti — un programma ancora in larga parte aperto.

Un secondo limite riguarda la risoluzione spaziale delle tecniche disponibili. Il predictive coding predice che messaggi ascendenti e discendenti abbiano proprietà laminari specifiche: gli errori di predizione sarebbero trasmessi da cellule piramidali superficiali negli strati II/III, le predizioni discendenti arriverebbero tramite fibre agli strati I e VI. Verificare queste predizioni richiede misurazioni con risoluzione laminare che solo di recente, con l'elettrofisiologia a ultra-alta densità nei primati e la risonanza a ultra-alto campo, stanno diventando parzialmente accessibili.

II. Precisione e attenzione: il test sperimentale

La teoria dell'attenzione come regolazione della precisione genera predizioni abbastanza specifiche da distinguerla da teorie concorrenti. Se l'attenzione corrisponde all'aumento della precisione Π degli errori di predizione al livello rilevante, si dovrebbe osservare che dirigere l'attenzione verso uno stimolo amplifichi le risposte neurali in modo proporzionale alla riduzione della varianza degli errori — non soltanto alla forza del segnale, ma al rapporto segnale-rumore.

Questo è compatibile con i dati sull'effetto di attenzione sulla componente P1 del potenziale visivo evocato, che mostra una amplificazione del 30-50% in risposta all'attenzione spaziale nei compiti di Posner. La predizione più specifica riguarda l'interazione tra attenzione e incertezza contestuale: se l'attenzione è stima della precisione, dovrebbe essere più intensa in contesti di elevata incertezza. Studi di Auksztulewicz e Friston e di Vossel e collaboratori

hanno trovato evidenze coerenti con questa predizione in compiti di rilevamento segnale con incertezza variabile, ma la letteratura rimane in fase di consolidamento.

III. Il problema del legame tra grandezze neurali e parametri variazionali

Un punto critico spesso trascurato riguarda il problema dell'identificazione: quali grandezze neurali misurabili corrispondono ai parametri del modello variazionale? Le equazioni del predictive coding operano su μ^(l) (credenze medie), ξ^(l) (errori di predizione ponderati), Π^(l) (precisioni). Ma cosa sono queste variabili in termini di attività neurale?

Le proposte più sviluppate identificano μ^(l) con il tasso di scarica sostenuto delle cellule piramidali profonde (strato V/VI), ξ^(l) con il tasso di scarica delle cellule piramidali superficiali (strato II/III), e Π^(l) con il guadagno sinaptico modulato dalla neurotrasmissione colinergica e dopaminergica. Queste identificazioni sono plausibili e parzialmente supportate da dati elettrofisiologici, ma rimangono approssimazioni: ci sono molte classi neuronali il cui ruolo nel framework non è ancora chiaramente compreso.

Il problema è aggravato dalla transizione tra il formalismo continuo e la neurofisiologia discreta. Il rate coding — che assume che l'informazione sia codificata nel tasso medio di scarica su finestre di decine di millisecondi — può perdere informazione contenuta nei pattern temporali precisi degli spikes. Se parte dell'elaborazione corticale utilizza temporal coding preciso, il formalismo del predictive coding gaussiano potrebbe sistematicamente sottocapturare quella componente.

IV. La teoria computazionale della mente di Fodor e le sue tensioni con il FEP

Jerry Fodor è stato il teorico più influente della mente computazionale nella seconda metà del Novecento, e il suo sistema intellettuale — articolato soprattutto in *The Language of Thought* (1975) e *The Modularity of Mind* (1983) — rimane il termine di confronto per qualsiasi teoria computazionale della cognizione. L'esame delle differenze tra la posizione fodoriana e il FEP rivela con chiarezza le assunzioni fondamentali di entrambe.

La teoria computazionale della mente di Fodor poggia su tre pilastri. Il primo è il Linguaggio del Pensiero: i processi cognitivi centrali sono computazioni su rappresentazioni con struttura sintattica analoga a quella del linguaggio naturale — rappresentazioni composizionali con costituenti e relazioni strutturali. Il secondo è il sistema classico: le operazioni cognitive rispettano la struttura sintattica delle rappresentazioni, garantendo la sistematicità e la produttività del pensiero. Il terzo è la modularità: il sistema cognitivo è composto da moduli di input (sistemi percettivi specializzati) e da processi centrali di ragionamento che operano sugli output dei moduli.

La tensione tra l'architettura fodoriana e il FEP è profonda e multipla.

Sul piano della struttura delle rappresentazioni, il FEP adotta rappresentazioni distribuite e continue — distribuzioni di probabilità nello spazio degli stati nascosti — piuttosto che simboli discreti con struttura sintattica. Nel predictive coding, il contenuto di una credenza è codificato dal valore medio μ e dalla covarianza Σ di una gaussiana. Questo rende la composizionalità molto più difficile da garantire. Fodor e Pylyshyn avevano identificato questo come il problema fondamentale del connessionismo (di cui il predictive coding è una variante sofisticata), e la sfida rimane: come derivare la sistematicità e la produttività del pensiero da rappresentazioni distribuite continue?

Sul piano della modularità, il FEP propone un'architettura radicalmente anti-fodoriana: un sistema gerarchico in cui ogni livello riceve input sia dal basso (segnali sensoriali) sia dall'alto (predizioni), e in cui i processi centrali di credenza sono in continua interazione con i processi periferici di elaborazione sensoriale. Questa penetrazione cognitiva — la capacità dei processi centrali di influenzare l'elaborazione percettiva — è precisamente ciò che Fodor negava per i moduli di input. Per Fodor, la penetrazione cognitiva era una minaccia epistemica: se la percezione è influenzata dalle credenze, come può fungere da base non circolare per la verifica delle credenze? Per il FEP, questa penetrazione è costitutiva della struttura dell'inferenza.

Sul fronte dell'olismo, Fodor era un critico severo della tesi che il contenuto di una credenza dipenda dal suo posto nell'intera rete di credenze del sistema, sostenendo che l'olismo rende la psicologia scientificamente intrattabile. Il FEP è costitutivamente olistico: il modello generativo è una distribuzione congiunta su tutti gli stati nascosti, e l'energia libera è un funzionale globale. Questo olismo non è accidentale: è una conseguenza necessaria dell'approccio bayesiano, in cui le credenze si aggiornano in modo coerente con l'intera struttura probabilistica del modello.

V. La critica della sistematicità e la risposta del FEP

L'obiezione della sistematicità — formulata da Fodor e Pylyshyn nel 1988 contro il connessionismo — merita una discussione separata. L'argomento è il seguente: il pensiero umano è sistematico in modo costitutivo, non contingente. Chiunque possa pensare «Giovanni ama Maria» può anche pensare «Maria ama Giovanni». Questa sistematicità richiede una spiegazione in termini di componenti condivisi. Il LoT la fornisce naturalmente; i modelli distribuiti non hanno le risorse per garantirla a priori.

Come risponde il FEP? La risposta più plausibile è che la sistematicità emerge dalla struttura fattorizzata del modello generativo. Se il modello fattorizza gli stati nascosti in variabili relativamente indipendenti — come avviene in un modello che apprende rappresentazioni disgiunte delle cause del mondo — allora le combinazioni di queste variabili sono sistematicamente rappresentabili. Formalmente, se:

$p(\psi) = p(\psi_\text{agente}) \cdot p(\psi_\text{paziente}) \cdot p(\psi_\text{relazione})$

allora le credenze su «Giovanni ama Maria» e «Maria ama Giovanni» differiscono soltanto per il valore di ψ_agente e ψ_paziente — strutture che condividono gli stessi costituenti con ruoli scambiati. La sistematicità è garantita dalla struttura fattorizzata del priore.

Questa risposta è promettente ma non definitiva. Resta la questione di come il cervello apprenda a fattorizzare il modello generativo in variabili corrispondenti a entità e relazioni rilevanti. I recenti progressi nei modelli generativi profondi — i transformer e i modelli a variabili latenti strutturate — suggeriscono che l'apprendimento di rappresentazioni composizionali da dati grezzi è possibile, ma i meccanismi biologici plausibili rimangono poco chiari.

VI. Il connessionismo e la sua eredità nel FEP

Il FEP può essere visto come la sintesi più matura della tradizione connessionista, ma una sintesi che ha assorbito le critiche fodoriane trasformandole in vincoli costruttivi.

Il connessionismo classico — il Parallel Distributed Processing di Rumelhart e McClelland — era un programma di apprendimento statistico: le reti imparano tramite backpropagation. Ma questo algoritmo era biologicamente inverosimile (richiede simmetria tra pesi in avanti e all'indietro non osservabile nel cervello), e l'architettura era determinata empiricamente piuttosto che derivata da principi.

Il FEP supera entrambe queste limitazioni. Sul fronte della plausibilità biologica, il predictive coding fornisce un algoritmo di apprendimento locale — la modifica sinaptica è guidata dagli errori di predizione locali, senza propagazione globale. Formalmente, la regola di apprendimento per i pesi w_{ij} tra l'unità j al livello l+1 e l'unità i al livello l è:

$$\Delta w_{ij} \propto \xi_i^{(l)} \cdot \mu_j^{(l+1)}$$

ossia il prodotto dell'errore di predizione all'unità i per il valore dell'unità generatrice j — una regola hebbiana modificata biologicamente plausibile. Millidge, Tschantz e Buckley hanno dimostrato che il predictive coding con questa regola locale converge allo stesso minimo del backpropagation in reti feed-forward, risolvendo un problema aperto sulla plausibilità biologica dell'apprendimento supervisionato.

Sul fronte del principio unificante, il FEP fornisce esattamente ciò che il connessionismo classico mancava: un principio normativo (minimizzazione dell'energia libera) da cui derivare sia l'architettura della rete sia la regola di apprendimento sia il comportamento dell'agente.

VII. Note critiche all'inflazionismo ontologico di Friston

Un tema ricorrente nelle critiche al programma fristoniano è l'inflazionismo ontologico: la tendenza a estendere il vocabolario del FEP ben al di là del dominio in cui è empiricamente

giustificato. Friston ha sostenuto in alcune pubblicazioni che qualsiasi sistema fisico con un confine di Markov stabile — incluse cellule, virus, persino sistemi chimici auto-organizzati — può essere descritto come sistema che minimizza l'energia libera e compie inferenza bayesiana. Da questa universalizzazione, alcuni collaboratori hanno tratto conclusioni prossime al panpsichismo.

La critica più acuta, formulata tra gli altri da Ramstead e Hohwy, distingue due affermazioni: (a) un sistema può essere descritto nel vocabolario dell'inferenza bayesiana; (b) un sistema compie effettivamente inferenza bayesiana. La prima è un'affermazione sulla prospettiva dell'osservatore — sulle descrizioni che un teorico può adottare. La seconda è un'affermazione sulla natura intrinseca del sistema. Il FEP giustifica soltanto la prima.

Quando Friston afferma che le cellule «credono» di trovarsi all'interno di corpi o che i batteri «inferiscono» la presenza di gradienti chimici, il vocabolario intenzionale è usato metaforicamente o letteralmente? Se metaforicamente, non c'è problema filosofico ma c'è rischio di confusione. Se letteralmente, si richiede una teoria dell'intenzionalità che spieghi in quali condizioni un processo fisico costituisce genuinamente una credenza — teoria che il FEP da solo non fornisce.

VIII. Realismo strutturale sulle distribuzioni di probabilità

Una nota critica finale riguarda il fondamento epistemologico del formalismo bayesiano. Il FEP assume che il cervello mantenga distribuzioni di probabilità sugli stati nascosti — ma cosa sono queste distribuzioni? Oggetti reali nel cervello, o strumenti matematici del teorico?

Se le distribuzioni sono reali, il FEP fa affermazioni empiriche verificabili sulla struttura dell'attività neurale. Se invece sono soltanto strumenti descrittivi, il FEP rischia di essere irrefutabile nel senso popperiano: qualsiasi sistema può essere descritto come minimizzatore di energia libera.

La posizione più difendibile è il realismo strutturale: il cervello implementa strutture che sono isomorfe alle distribuzioni di probabilità del formalismo, senza assumere che le distribuzioni esistano letteralmente come oggetti nel cervello. Questa posizione è analoga al realismo strutturale in filosofia della fisica — la tesi che le teorie fisiche catturino la struttura della realtà senza che si debba ipostatizzare ogni entità teorica. È epistemicamente umile ma non deflazionista: riconosce il potere esplicativo della struttura matematica riconoscendone al tempo stesso il carattere di idealizzazione.

Tre Specchi per il Principio di Energia Libera

I. La logica del confronto

I tre territori che si esplorano in questa parte non sono appendici tematiche: sono specchi. Ciascuno riflette il FEP da un angolo diverso, e in ciascun riflesso appaiono sia le zone di convergenza sia — più utilmente — le zone di opacità, quelle dove la teoria non si vede o si deforma. Il metodo non sarà enciclopedico ma dialettico: si cercherà ogni volta il punto di massima tensione tra il FEP e la teoria in questione, perché è lì che entrambe rivelano le proprie assunzioni più profonde.

Il filo che li connette è la domanda lasciata aperta dalla Parte Quinta: cosa trasforma la preferenza strutturale in preferenza vissuta? Metzinger la affronta chiedendo quali condizioni rendono un modello di sé trasparente all'agente che lo possiede. Dehaene e Baars la affrontano chiedendo quale architettura globale rende un contenuto neurale accessibile alla coscienza. Piaget — e la psicologia dello sviluppo che lo prolonga — la affronta chiedendo come un sistema privo di modello del mondo ne costruisce uno a partire dall'esperienza grezza. Nessuno dei tre risponde alla domanda come è stata posta, ma ciascuno la arricchisce in un modo che il FEP da solo non avrebbe potuto.

II. Metzinger e la trasparenza del modello di sé

La *Self-Model Theory of Subjectivity* (SMT) di Thomas Metzinger, sviluppata con sistematicità filosofica impressionante in *Being No One* (2003) e successivamente in *The Ego Tunnel* (2009), condivide con il FEP l'intuizione centrale che il sé non sia un'entità ma un processo — più precisamente, il contenuto di un modello generativo che il sistema nervoso mantiene su se stesso. Ma da questa intuizione condivisa Metzinger trae conseguenze che il FEP non cattura e che illuminano con precisione dove l'abisso esplicativo rimane aperto.

Il concetto cardine della SMT è quello di *phenomenal self-model* (PSM): la rappresentazione neurale che il sistema nervoso costruisce di se stesso in quanto organismo — della propria postura, del proprio stato interoceptivo, della propria identità psicologica. Questa rappresentazione ha una proprietà peculiare che Metzinger chiama *trasparenza fenomenica*: il soggetto non la riconosce come una rappresentazione. Quando percepiamo la nostra mano come nostra, non percepiamo una rappresentazione della mano: percepiamo la mano direttamente, come se il modello fosse una finestra trasparente sul sé piuttosto che uno specchio opaco che lo riflette. Questa trasparenza non è un'illusione cognitiva corretta per mezzo della riflessione: è la struttura di default dell'esperienza soggettiva nella sua forma ordinaria.

Il punto di connessione con il FEP è immediato: se il sé è il contenuto del modello generativo del sistema su se stesso, allora la trasparenza fenomenica dovrebbe corrispondere a qualche proprietà del processo di minimizzazione dell'energia libera. La domanda è quale. Metzinger suggerisce che la trasparenza emerga quando il sistema non ha accesso alle rappresentazioni di ordine superiore che raffigurano il PSM *come* rappresentazione — quando la meta-rappresentazione è assente o soppressa. In condizioni normali, il cervello elabora le informazioni somatiche e psicologiche troppo velocemente perché i livelli superiori della

gerarchia possano "vedere" che il risultato è una costruzione: il tempo di elaborazione è inferiore alla soglia di latenza della meta-rappresentazione.

Nel vocabolario del FEP, questa idea si traduce nel modo seguente. Il PSM è il valore medio μ^(sé) del livello più alto della gerarchia che riguarda il sistema stesso. La trasparenza fenomenica corrisponde alla condizione in cui l'errore di predizione relativo a μ^(sé) è soppresso — quando le predizioni auto-relative sono costantemente soddisfatte con alta precisione, tanto da non generare segnali di errore che risalgano verso ulteriori livelli di meta-rappresentazione. In assenza di errori di predizione significativi sul PSM, non c'è segnale che il modello del sé stia *modellando* qualcosa: il modello si rende invisibile.

Questa lettura permette di fare una predizione verificabile che Metzinger non aveva formulato esplicitamente: la trasparenza fenomenica dovrebbe diminuire — e la natura costruttiva del PSM dovrebbe diventare più evidente — nelle condizioni in cui la precisione dei segnali sensoriali che alimentano il PSM viene artificialmente ridotta o perturbata. Questo è precisamente ciò che si osserva nelle esperienze di depersonalizzazione, nelle fluttuazioni del senso di agency sotto farmaci dissociativi come la ketamina (che blocca i recettori NMDA e pertuba la trasmissione che sostiene la precisione dei segnali predittivi), e nell'esperienza della mano di gomma — il classico paradigma in cui la sincronia sensoriale artificiosa porta il soggetto ad attribuire la proprietà corporea a un oggetto inerte. In tutti questi casi, la perturbazione della precisione dei segnali del PSM genera errori di predizione che risalgono la gerarchia, rendendo il modello di sé *opaco* — visibile come costruzione — in modo fenomenicamente saliente.

La divergenza profonda tra la SMT e il FEP emerge però su un punto preciso. Per Metzinger, la trasparenza fenomenica è una proprietà costitutiva dell'esperienza soggettiva: non c'è esperienza in prima persona senza che il PSM sia trasparente, senza che il sistema non si accorga che sta modellando se stesso. Questa tesi ha un corollario sorprendente: i sistemi che non hanno un PSM trasparente — che elaborano informazioni su se stessi senza la trasparenza costitutiva — non hanno esperienza soggettiva nel senso pieno. Il problema difficile, in questa prospettiva, si trasforma nel problema della trasparenza: spiegare non perché ci sia esperienza ma perché certi modelli siano opachi ai propri processi di modellazione.

Il FEP può *descrivere* le condizioni di precisione sotto cui la trasparenza emerge, ma non può *derivare* che dalla soppressione degli errori di predizione sul PSM emerga la trasparenza fenomenica. È una condizione necessaria, non sufficiente. L'implicazione che Metzinger vuole trarre — che la trasparenza sia identica alla fenomenicità, che non ci sia nulla oltre la struttura funzionale del PSM trasparente — è, ancora una volta, una tesi illusionista mascherata da tesi strutturale. Il critico direbbe: un sistema può avere un PSM altamente trasparente, con errori di predizione sul sé completamente soppressi, e tuttavia rimanere uno zombie funzionale. La trasparenza è una proprietà del processo di rappresentazione, non dell'esperienza di quel processo.

Ciò che la SMT aggiunge al FEP non è la soluzione di questo residuo ma la sua precisione topografica. La trasparenza identifica il luogo esatto nella gerarchia inferenziale in cui il problema difficile si installa: non nella percezione del mondo esterno, non nelle rappresentazioni di ordine basso, ma nella struttura del PSM al livello più alto della gerarchia

auto-riferita. Sapere dove il problema vive è diverso dal risolverlo, ma è un progresso reale rispetto alla vaga affermazione che l'esperienza emerge dalla complessità neurale.

III. Il Global Workspace e la questione dell'accesso cosciente

La teoria del Global Workspace, proposta da Bernard Baars nei tardi anni '80 come modello cognitivo e formalizzata computazionalmente da Stanislas Dehaene, Jean-Pierre Changeux e collaboratori nel decennio successivo, parte da una domanda diversa da quella del FEP: non quale principio governa l'elaborazione cognitiva in generale, ma quale meccanismo distingue i contenuti neurali che raggiungono la coscienza da quelli che rimangono inconsci.

La risposta di Baars era metaforica ma fertile: la coscienza è come un teatro in cui molti processi specializzati e inconsci competono per accedere a un palcoscenico globale — il *global workspace* — da cui l'informazione viene trasmessa in broadcast a tutti i sistemi specializzati contemporaneamente. I contenuti che raggiungono il palcoscenico diventano coscienti; quelli che restano nelle aree locali rimangono inconsci. La versione computazionale di Dehaene e Changeux traduce questa metafora in termini neurali precisi: il global workspace corrisponde a una rete di neuroni piramidali a lungo range nelle regioni prefrontali e parietali, con assoni altamente mielinizzati che permettono la trasmissione rapida di informazioni attraverso il cervello. Quando un contenuto viene amplificato abbastanza da innescare un'ignizione (*ignition*) di questa rete — una transizione non lineare tra uno stato locale e uno stato globale di attivazione sincronizzata — diventa cosciente.

Le predizioni empiriche della GWT sono specifiche e verificate con una base sperimentale oggi molto robusta. Nei paradigmi di masking, in cui uno stimolo brevissimo viene reso invisibile da uno stimolo successivo, si osserva una dissociazione netta tra l'elaborazione neurale locale (che persiste anche per stimoli mascherati) e l'ignizione globale (che avviene solo per stimoli visibili). I dati di EEG e MEG mostrano un marker elettrofisiologico caratteristico della coscienza — la *P3b*, un'onda positiva tardiva attorno ai 300-500 ms — che si correla con l'ignizione della rete prefrontale-parietale, non con l'elaborazione sensoriale precoce. La GWT è al momento la teoria della coscienza con il maggiore supporto empirico diretto per quanto riguarda la coscienza di accesso.

Il confronto con il FEP rivela una complementarità strutturale profonda e alcune tensioni altrettanto profonde. La complementarità è la seguente: la GWT descrive quale contenuto diventa cosciente — quello che raggiunge il global workspace — mentre il FEP descrive come i contenuti vengono generati e selezionati attraverso la minimizzazione dell'energia libera. I due livelli di descrizione sono compatibili e si integrano: si può proporre che il global workspace sia il meccanismo mediante cui i segnali di errore di predizione con alta precisione vengono *amplificati globalmente* e resi disponibili per l'aggiornamento dei modelli generativi di alto livello. L'ignizione del workspace sarebbe, in questa lettura, la transizione dinamica attraverso cui un errore di predizione localmente significativo viene elevato a segnale globale che forza l'aggiornamento coerente di tutta la gerarchia.

Formalmente, si può immaginare che l'ignizione del global workspace corrisponda a una biforcazione nel sistema dinamico degli stati neurali: al di sotto di una soglia di precisione Π^*,

gli errori di predizione rimangono confinati nell'elaborazione locale; al di sopra di quella soglia, l'amplificazione non lineare produce una transizione a uno stato di attivazione globale che coinvolge le proiezioni prefrontali a lungo range. La soglia Π^* dipenderebbe dall'architettura della rete prefrontale-parietale e dalla neurotrasmissione modulatoria (in particolare la noradrenalina, che aumenta il rapporto segnale-rumore nelle reti prefrontali). In questo quadro, l'attenzione — la regolazione della precisione Π nel FEP — è il meccanismo che controlla quali errori di predizione raggiungono la soglia di ignizione e quindi la coscienza di accesso.

Questa integrazione è elegante, ma nasconde una tensione che merita esposizione onesta. La GWT, nelle versioni più ambiziose di Dehaene, aspira a spiegare non soltanto la coscienza di accesso ma anche la coscienza fenomenica: l'ignizione globale non sarebbe soltanto il meccanismo dell'accessibilità cognitiva ma la base sufficiente per l'esperienza soggettiva. Questa ambizione è contestata da Ned Block, che distingue le due nozioni con precisione chirurgica, e da Chalmers, che osserva che anche se si identificasse con certezza il correlato neurale della coscienza di accesso, resterebbe intatta la domanda di perché questo correlato sia accompagnato da esperienza qualitativa.

Il FEP condivide con la GWT questa limitazione: entrambe le teorie descrivono con crescente precisione il substrato funzionale e neurale della coscienza di accesso, e entrambe hanno difficoltà a fare il passo dalla struttura all'esperienza. La convergenza tra FEP e GWT su questo punto non è casuale: entrambe appartengono alla classe delle teorie funzionali della coscienza, e il problema difficile è precisamente il problema del residuo che le teorie funzionali lasciano irrisolto.

Dove però la GWT aggiunge qualcosa che il FEP da solo non ha è nella specificità della domanda sul *quando* e *quale*: la GWT specifica le condizioni dinamiche precise in cui un contenuto passa dalla zona inconscia a quella cosciente — il momento dell'ignizione, la soglia, la geometria della biforcazione. Il FEP specifica il *perché* i contenuti vengono selezionati nella gerarchia, ma non il meccanismo della transizione alla coscienza di accesso. L'integrazione delle due teorie potrebbe produrre qualcosa di più preciso di ciascuna presa separatamente: una teoria in cui la selezione dei contenuti da parte del FEP (tramite la regolazione della precisione) e il meccanismo di diffusione globale della GWT (tramite l'ignizione prefrontale) si specificano vicendevolmente, producendo predizioni quantitative su quali perturbazioni del sistema di precisione alterano la soglia di ignizione e quindi alterano il contenuto della coscienza di accesso.

Vi è però un punto in cui la GWT resiste all'assorbimento nel FEP e lo fa in modo illuminante. Il global workspace è essenzialmente un meccanismo di *trasmissione* — un sistema che diffonde informazioni già elaborate a tutto il cervello. Non è un sistema che genera ipotesi sul mondo: è un sistema che seleziona quali ipotesi diventano globalmente accessibili. Il predictive coding, al contrario, è essenzialmente un meccanismo di *generazione* — un sistema che produce ipotesi e le confronta con l'evidenza. La tensione tra generazione e selezione è una tensione reale: la GWT suggerisce che la coscienza sia un fenomeno di selezione e trasmissione, mentre il FEP la vede come un fenomeno di generazione e minimizzazione. Riconciliare queste due prospettive senza perdere la specificità di ciascuna rimane un problema aperto.

IV. L'ontogenesi del modello generativo: Piaget rovesciato

Il terzo territorio è forse il più fondamentale, perché tocca una presupposizione silenziosa che attraversa tutto il formalismo del FEP: che il sistema disponga già di un modello generativo sul mondo. Tutta la struttura matematica della minimizzazione dell'energia libera — le distribuzioni a priori, le verosimiglianze, la gerarchia dei modelli — presuppone che i parametri del modello siano già specificati. Ma da dove vengono? Come li apprende un neonato?

La domanda non è tecnica: è fondativa. Un sistema privo di modello generativo non ha distribuzioni a priori da aggiornare, non ha predizioni da confrontare con l'evidenza, non ha errori di predizione da minimizzare. Non è un minimizzatore di energia libera nel senso operativo del termine: è un sistema che deve *diventare* un minimizzatore di energia libera costruendo, attraverso l'esperienza, il modello che poi minimizzerà. Il paradosso è quello del bootstrapping: per apprendere un modello del mondo bisogna già avere un modello del mondo sufficientemente ricco da orientare l'apprendimento.

Piaget aveva affrontato una versione di questo paradosso con la sua teoria degli schemi e dei processi di assimilazione e accomodamento. Per Piaget, il neonato non è una tabula rasa ma un sistema dotato di schemi senso-motori elementari — i riflessi di suzione, prensione, orientamento — che costituiscono il punto di partenza dell'apprendimento. L'assimilazione incorpora nuove esperienze negli schemi esistenti; l'accomodamento modifica gli schemi quando le esperienze non vi si adattano. Lo sviluppo cognitivo è la costruzione progressiva di strutture conoscitive sempre più astratte attraverso l'interazione con il mondo.

La lettura di Piaget nel vocabolario del FEP è suggestiva ma asimmetrica. L'assimilazione è la minimizzazione dell'energia libera attraverso l'aggiornamento percettivo: nuove esperienze vengono interpretate alla luce del modello esistente, che rimane invariato. L'accomodamento è la minimizzazione dell'energia libera attraverso l'aggiornamento del modello: la struttura delle distribuzioni a priori e delle verosimiglianze si modifica per rendere conto di esperienze che il modello corrente non riesce a spiegare. La distinzione piagetiana tra i due processi corrisponde alla distinzione del FEP tra aggiornamento delle variabili latenti (inferenza percettiva) e aggiornamento dei parametri del modello (apprendimento).

Ma il FEP inverte il privilegio esplicativo rispetto a Piaget in un modo che merita attenzione. Piaget partiva dall'azione: i primi schemi sono senso-motori, e la cognizione astratta emerge per internalizzazione progressiva degli schemi motori. Il FEP, almeno nella sua versione standard, parte dalla percezione: il sistema inferisce gli stati nascosti del mondo e poi agisce per ridurre la sorpresa. Questa asimmetria non è innocua. Per un neonato, i cui sistemi sensoriali sono grossolani e i cui sistemi motori sono ancora in via di mielinizzazione, l'azione è più affidabile della percezione: i pattern motori ritmici — suzione, movimento ciclico degli arti — producono conseguenze sensoriali prevedibili molto prima che il sistema visivo raggiunga la maturità necessaria per costruire priori affidabili sul mondo. L'azione non segue la percezione: la precede e la struttura.

Questa inversione ha una formalizzazione nel FEP che non è stata sviluppata abbastanza esplicitamente nella letteratura. Se si considera che il neonato inizia con priori su *policy* (schemi motori innati) piuttosto che con priori su *stati del mondo*, allora il processo di costruzione del modello generativo comincia dall'interno: il sistema apprende innanzitutto la struttura delle conseguenze sensoriali delle proprie azioni — impara p(o | a), la verosimiglianza dell'input sensoriale dato un certo schema motorio — prima di apprendere la struttura del mondo indipendente dall'azione. Il modello del mondo emerge come inferenza inversa sulle conseguenze sistematiche dell'azione: gli stati nascosti ψ vengono costruiti come le cause ipotizzate che spiegano perché certe azioni producono sistematicamente certi pattern sensoriali.

In termini formali, si può pensare al modello generativo neonatale come inizialmente fattorizzato nel modo seguente:

$$p_0(o, \psi, a) \approx p(o \mid a) \cdot p(a) \cdot p_0(\psi)$$

dove p(o | a) è una distribuzione relativamente specifica (gli schemi motori innati producono conseguenze sensoriali abbastanza stabili), p(a) è il priore innato sulle politiche motorie (i riflessi), e $p_0(\psi)$ è una distribuzione molto piatta sugli stati del mondo (pochissima conoscenza a priori del mondo esterno). L'apprendimento ontogenetico consiste nel ridistribuire la specificità: dalla specificità su p(o | a) verso la specificità su p(o | ψ) e p(ψ), ossia dalla conoscenza delle conseguenze delle proprie azioni verso la conoscenza della struttura del mondo indipendente dall'azione.

Questa rilettura dello sviluppo cognitivo nel vocabolario del FEP produce un corollario importante per la psicopatologia. Se lo sviluppo normale procede costruendo progressivamente una struttura ψ del mondo a partire dalla struttura a delle azioni, allora le perturbazioni precoci del sistema motorio — come quelle che si osservano in certi profili di rischio per lo spettro autistico, dove i pattern motori ritmici neonatali sono atipici — dovrebbero alterare non soltanto lo sviluppo motorio ma la struttura stessa del modello generativo del mondo che su quei pattern viene costruito. Questa predizione è coerente con osservazioni cliniche sull'atipicità dei movimenti precoci nei bambini che riceveranno successivamente una diagnosi di autismo, e con la proposta di Lawson e Friston discussa nella Parte Sesta — ma la radica in una genealogia ontogenetica che la proposta originaria non esplicitava.

V. Il nodo comune: la temporalità dell'inferenza

I tre territori esplorati convergono su un tema che il FEP nella sua versione standard tratta in modo insufficiente: la temporalità dell'inferenza. Il PSM di Metzinger è un modello che si costruisce nel tempo e può collassare o dissolversi; l'ignizione del global workspace è un evento temporalmente localizzato, una transizione di fase; lo sviluppo cognitivo piagetiano è un processo di costruzione che si dispiega su anni. Il FEP, nella sua formulazione stazionaria, tratta il modello generativo come dato e descrive la dinamica dell'inferenza su di esso: ma non ha una teoria della genesi temporale del modello stesso.

Questa lacuna è più che tecnica. Il problema difficile della coscienza ha una dimensione temporale spesso trascurata: non soltanto c'è qualcosa che è come essere un sistema, ma c'è qualcosa che è come *essere un sistema nel tempo* — l'esperienza ha una struttura temporale intrinseca, un flusso, una ritenzione del passato e una protensione verso il futuro che Husserl aveva analizzato con straordinaria finezza nella sua fenomenologia della coscienza del tempo interno. Il FEP, con la sua struttura di politiche e di traiettorie future attese, cattura in parte questa dimensione — la protensione è formalizzata nell'energia libera attesa $G(\pi)$ — ma la ritenzione, la struttura della memoria episodica come fondamento dell'identità narrativa del sé, rimane largamente fuori dalla sua portata.

Qui Metzinger, la GWT e Piaget convergono in modo inatteso. Metzinger ha sostenuto che il PSM include una componente temporale — il *phenomenal self as temporal* — che codifica la continuità del sé attraverso il tempo; la GWT, nelle recenti elaborazioni di Dehaene, include una distinzione tra contenuti coscienti di accesso e contenuti dell'autobiografia (la *conscious access to one's own past*) che richiede un'architettura di memoria distinta da quella del workspace attuale; e Piaget aveva mostrato che la permanenza dell'oggetto — la capacità di mantenere la rappresentazione di un oggetto anche quando non è percepito — è una conquista tardiva dello sviluppo, che emerge attorno ai nove mesi, e rappresenta la costruzione della prima struttura di modello del mondo che ha spessore temporale.

In ciascuno di questi casi, il FEP ha le risorse formali per affrontare queste questioni — la struttura delle politiche come distribuzioni su traiettorie, il modello generativo come processo markoviano nascosto con memoria — ma non le ha ancora integrate in modo che faccia giustizia alla ricchezza delle osservazioni fenomenologiche, neuropsicologiche e sviluppiste. Questa integrazione è forse il programma di ricerca più promettente che il confronto tra il FEP e questi tre territori indica: non un'estensione dell'architettura matematica, che è già sufficientemente potente, ma un'applicazione sistematica di quella architettura alla struttura temporale dell'esperienza — alla costruzione del sé nel tempo, alla genesi della continuità cosciente, all'ontogenesi del modello generativo come processo che dura una vita intera.

La domanda con cui si chiude questa parte è pertanto la stessa con cui si chiudeva la quinta, ma formulata in modo più preciso grazie al percorso che la precede: non soltanto «qual è la metrica intrinseca degli stati interni di un sistema minimizzante l'energia libera?», ma «come questa metrica si costruisce nel tempo, e qual è la struttura temporale dell'esperienza che emerge da questo processo di costruzione?» Le due domande non si rispondono l'una con l'altra, ma insieme delimitano con precisione il territorio in cui una teoria completa della mente dovrà muoversi.

Appendici Tecniche e Bibliografia

Appendice A — Derivazione completa della disuguaglianza variazionale

Si fornisce qui la derivazione rigorosa della relazione fondamentale $F \geq -\ln p(o)$, seguendo la notazione adottata nel testo.

A.1 Punto di partenza

Dato un modello generativo $p(o, \psi)$ e una distribuzione di riconoscimento $q(\psi)$, si definisce l'energia libera variazionale:

$$F[q, o] \equiv \int q(\psi) \ln[q(\psi) / p(o, \psi)] \, d\psi$$

Si vuole mostrare che $F[q, o] \geq -\ln p(o)$ con uguaglianza se e solo se $q(\psi) = p(\psi \mid o)$.

A.2 Dimostrazione

Si aggiunge e sottrae $\ln p(o)$ nell'integrando:

$$F = \int q(\psi) [\ln q(\psi) - \ln p(o, \psi)] \, d\psi$$

$$= \int q(\psi) [\ln q(\psi) - \ln p(\psi \mid o) - \ln p(o)] \, d\psi$$

$$= \int q(\psi) [\ln q(\psi) - \ln p(\psi \mid o)] \, d\psi - \ln p(o) \cdot \int q(\psi) \, d\psi$$

Poiché q è una densità di probabilità, $\int q(\psi) \, d\psi = 1$, quindi:

$$F = KL[q(\psi) \;/\!/\; p(\psi \mid o)] - \ln p(o)$$

Per la disuguaglianza di Jensen applicata alla funzione convessa $-\ln(\cdot)$ (o equivalentemente per la non-negatività della divergenza KL, che si dimostra mediante la disuguaglianza di Gibbs):

$$KL[q \;/\!/\; p] = \int q(\psi) \ln[q(\psi)/p(\psi|o)] \, d\psi \geq 0$$

con uguaglianza se e solo se $q(\psi) = p(\psi \mid o)$ quasi ovunque. Pertanto:

$$F = KL[q(\psi) \;/\!/\; p(\psi \mid o)] - \ln p(o) \geq -\ln p(o)$$

che è quanto si voleva dimostrare. Il termine $-\ln p(o)$ è la sorpresa dell'osservazione, e F ne costituisce un limite superiore stretto, avvicinabile fino alla coincidenza quando q collassa sulla vera posteriore. □

A.3 Corollario: F come funzionale da minimizzare

Dalla decomposizione $F = KL[q \;/\!/\; p(\cdot \mid o)] - \ln p(o)$ si vede che:

$$\min_q F[q, o] = -\ln p(o)$$

poiché il minimo della KL è zero. L'inferenza variazionale consiste pertanto nel trovare la $\hat{q}$ che realizza (o approssima) questo minimo all'interno della famiglia distribuzionale prescelta.

Appendice B — Derivazione del gradiente naturale e connessione con la metrica di Fisher

B.1 Il gradiente ordinario e i suoi limiti

L'aggiornamento delle credenze medie nella discesa del gradiente ordinaria è:

$$\mu(t+1) = \mu(t) - \alpha \cdot \partial F/\partial \mu$$

dove α è il passo di aggiornamento. Questo aggiornamento è dipendente dalla parametrizzazione: se si reparametrizzano le credenze con una trasformazione nonlineare $\varphi = h(\mu)$, il gradiente cambia forma in modo non covariante, e la traiettoria di ottimizzazione nello spazio delle distribuzioni dipende dalla scelta arbitraria dei parametri.

B.2 La metrica di Fisher come metrica naturale

Lo spazio delle distribuzioni di probabilità ha una struttura geometrica naturale — la geometria dell'informazione sviluppata da Amari — in cui la distanza tra due distribuzioni vicine $p(x\,;\theta)$ e $p(x\,;\theta + d\theta)$ è misurata dalla divergenza KL infinitesima:

$$ds^2 = d\theta^T\, g(\theta)\, d\theta$$

dove la matrice di informazione di Fisher g_{ij}(θ) è:

$$g_{ij}(\theta) = \mathbb{E}_{p(x;\theta)}[\partial_i \ln p(x;\theta) \cdot \partial_j \ln p(x;\theta)]$$

$$= -\mathbb{E}_{p(x;\theta)}[\partial_i\, \partial_j \ln p(x;\theta)]$$

(la seconda uguaglianza vale sotto condizioni di regolarità standard).

B.3 Il gradiente naturale

Il gradiente naturale è la direzione di massima discesa dell'energia libera nello spazio delle distribuzioni dotato della metrica di Fisher:

$$\nabla F = g(\theta)^{-1} \cdot \nabla F$$

L'aggiornamento del gradiente naturale è quindi:

$$\theta(t+1) = \theta(t) - \alpha \cdot g(\theta)^{-1} \cdot \partial F/\partial \theta$$

Questo aggiornamento è invariante rispetto a reparametrizzazioni: se si trasformano i parametri, il gradiente naturale si trasforma covariamente e la traiettoria di ottimizzazione nello spazio delle distribuzioni rimane la stessa.

B.4 Connessione con il predictive coding

Per una distribuzione gaussiana $q(\psi) = \mathcal{N}(\psi ; \mu, \Sigma)$, i parametri naturali sono $\eta_1 = \Sigma^{-1} \mu$ (media naturale) e $\eta_2 = -\frac{1}{2} \Sigma^{-1}$ (precisione). La matrice di Fisher in questi parametri è:

$$g(\eta) = \mathbb{E}_q[(\partial_\eta \ln q)(\partial_\eta \ln q)^T]$$

e si può mostrare che per la famiglia gaussiana il gradiente naturale rispetto alla media μ coincide con:

$$\tilde{\nabla}_\mu F = \Sigma \cdot \partial F/\partial \mu = \Sigma \cdot \xi$$

dove ξ è il vettore degli errori di predizione ponderati dalla precisione. Le equazioni di aggiornamento del predictive coding sono pertanto interpretabili come implementazioni approssimate del gradiente naturale nello spazio delle gaussiane — una proprietà che spiega la loro efficienza computazionale rispetto al gradiente ordinario.

Appendice C — L'energia libera attesa: derivazione della decomposizione rischio/ambiguità

C.1 Definizione

L'energia libera attesa per una politica π che genera stati futuri (o_τ, ψ_τ) con $\tau > t$ è:

$$G(\pi) = \mathbb{E}_{q(o_\tau, \psi_\tau \mid \pi)}[\ln q(\psi_\tau \mid \pi) - \ln p(o_\tau, \psi_\tau)]$$

C.2 Fattorizzazione del modello generativo futuro

Il modello generativo fattorizza come:

$$p(o_\tau, \psi_\tau) = p(o_\tau \mid \psi_\tau) \cdot p(\psi_\tau)$$

dove $p(\psi_\tau)$ codifica le preferenze dell'agente sugli stati futuri (lo stato a bassa sorpresa che il sistema cerca di raggiungere). La distribuzione variazionale fattorizza analogamente:

$$q(o_\tau, \psi_\tau \mid \pi) = p(o_\tau \mid \psi_\tau) \cdot q(\psi_\tau \mid \pi)$$

dove si assume che la verosimiglianza sia condivisa tra distribuzione variazionale e modello generativo.

C.3 Derivazione

Sostituendo nella definizione di G:

$$G(\pi) = \mathbb{E}_{q}[\ln q(\psi_\tau \mid \pi) - \ln p(o_\tau \mid \psi_\tau) - \ln p(\psi_\tau)]$$

Il secondo termine si può scrivere come:

$$-\mathbb{E}_{q}[\ln p(o_\tau \mid \psi_\tau)] = \mathbb{E}_{q(\psi_\tau|\pi)}[H[p(o_\tau \mid \psi_\tau)]] - \mathbb{E}_{q(\psi_\tau|\pi)}[\mathbb{E}_{p(o_\tau|\psi_\tau)}[\ln p(o_\tau \mid \psi_\tau)]]$$

Si definisce l'ambiguità come:

$$\text{Ambiguità}(\pi) \equiv \mathbb{E}_{q(\psi_\tau|\pi)}[H[p(o_\tau \mid \psi_\tau)]]$$

ossia l'incertezza media sulle osservazioni future data la stima degli stati nascosti futuri.

Raccogliendo i termini rimanenti:

$$\mathbb{E}_{q}[\ln q(\psi_\tau|\pi) - \ln p(\psi_\tau)] - \mathbb{E}_{q}[\mathbb{E}_{p(o_\tau|\psi_\tau)}[\ln p(o_\tau|\psi_\tau)]]$$

Il primo termine è:

$$KL[q(\psi_\tau|\pi) \;//\; p(\psi_\tau)] = \text{Rischio}(\pi)$$

che misura la divergenza tra la distribuzione degli stati futuri sotto la politica π e la distribuzione preferita degli stati futuri.

Il secondo termine è:

$$\mathbb{E}_{q(\psi_\tau|\pi)}[\mathbb{E}_{p(o_\tau|\psi_\tau)}[\ln p(o_\tau|\psi_\tau)]] = -\mathbb{E}_{q(\psi_\tau|\pi)}[H[p(o_\tau|\psi_\tau)]] + C$$

dove C è una costante rispetto a π. Quindi, a meno di costanti:

$$G(\pi) = KL[q(\psi_\tau|\pi) \;//\; p(\psi_\tau)] + \mathbb{E}_{q(\psi_\tau|\pi)}[H[p(o_\tau|\psi_\tau)]]$$

$G(\pi) = \text{Rischio}(\pi) + \text{Ambiguità}(\pi)$

La politica ottima minimizza la somma di rischio (divergenza dalle preferenze) e ambiguità (incertezza epistemica residua). Un agente che minimizza G è sia strumentalmente motivato (cerca gli stati preferiti) sia epistemicamente motivato (riduce l'incertezza sul mondo). □

Appendice D — Il confine di Markov nei sistemi di Langevin: derivazione della stazionarietà

D.1 Il processo di Langevin

Si consideri un sistema con variabili di stato $z = (\eta, s, a, \mu)$ che evolve secondo il processo di Langevin overdampito:

$$dz/dt = f(z) + \omega, \quad \omega \sim \mathcal{N}(0, 2\Gamma)$$

dove $f(z)$ è la funzione di deriva e Γ è la matrice di diffusione. Si assume che il sistema ammetta una distribuzione stazionaria $p^*(z)$ tale che $dp^*/dt = 0$.

D.2 La struttura del confine di Markov

Il confine di Markov richiede le condizioni di indipendenza condizionale:

$\partial f_\eta / \partial \mu = 0$ (gli stati esterni non dipendono direttamente dagli stati interni)
$\partial f_\mu / \partial \eta = 0$ (gli stati interni non dipendono direttamente dagli stati esterni)

Sotto queste condizioni, la densità stazionaria soddisfa la relazione di fattorizzazione condizionale:

$$p^*(\eta, \mu \mid b) = p^*(\eta \mid b) \cdot p^*(\mu \mid b)$$

ossia, condizionati al confine $b = (s, a)$, gli stati interni e gli stati esterni sono statisticamente indipendenti.

D.3 L'energia libera come funzione di Lyapunov

La densità stazionaria $p^*(z)$ soddisfa l'equazione di Fokker-Planck stazionaria:

$$0 = -\nabla \cdot (f\, p^*) + \nabla \cdot (\Gamma\, \nabla p^*)$$

Si può mostrare che, sotto condizioni di bilancio dettagliato generalizzato, la soluzione ha la forma:

$$p^*(z) \propto \exp(-\Psi(z))$$

dove $\Psi(z)$ è il potenziale del sistema. Per gli stati interni condizionati al confine, questo diventa:

$$p^*(\mu \mid b) \propto \exp(-F(\mu, b))$$

dove $F(\mu, b)$ è identificato con l'energia libera variazionale degli stati interni dato il confine. Gli stati interni di un sistema con confine di Markov stabile si distribuiscono, nella stazionarietà, come la distribuzione di Boltzmann di un sistema che minimizza l'energia libera — il risultato fondamentale che giustifica l'interpretazione inferenziale degli stati interni. □

Appendice E — La metrica di Fisher-Rao e la congettura sulla geometria fenomenica

E.1 La metrica di Fisher-Rao su famiglie parametriche

Sia $\{p(x\,;\theta) : \theta \in \Theta\}$ una famiglia di distribuzioni di probabilità parametrizzata da $\theta \in \mathbb{R}^d$. La metrica di Fisher-Rao $g(\theta)$ è il tensore metrico su Θ definito da:

$$g_{ij}(\theta) = \int p(x\,;\theta)\,[\partial_i \ln p(x\,;\theta)][\partial_j \ln p(x\,;\theta)]\,dx$$

$$= \mathbb{E}_{p(x;\theta)}[\partial_i \ln p \cdot \partial_j \ln p]$$

Questo tensore è simmetrico e, sotto condizioni di regolarità, positivo definito. Definisce una struttura riemanniana sullo spazio dei parametri.

E.2 La distanza geodetica

La distanza di Fisher-Rao tra due distribuzioni $p(x\,;\theta_1)$ e $p(x\,;\theta_2)$ è la lunghezza della geodetica che le connette:

$$d_{FR}(\theta_1, \theta_2) = \min_{\gamma : \gamma(0)=\theta_1,\, \gamma(1)=\theta_2} \int_0^1 \sqrt{\dot\gamma(t)^T\, g(\gamma(t))\, \dot\gamma(t)}\,dt$$

Per famiglie esponenziali, questa distanza ha espressioni analitiche chiuse. Per distribuzioni gaussiane univariate $\mathcal{N}(\mu, \sigma^2)$:

$$d_{FR}(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) = \sqrt{2}\cdot \operatorname{arcosh}\left(1 + \frac{(\mu_1-\mu_2)^2 + 2(\sigma_1-\sigma_2)^2}{2\sigma_1\sigma_2}\right)$$

dove arcosh è il coseno iperbolico inverso. Per distribuzioni gaussiane multivariate $\mathcal{N}(\mu, \Sigma)$:

$$d_{FR}^2(\theta_1, \theta_2) \approx (\mu_1-\mu_2)^T \Sigma^{-1} (\mu_1-\mu_2) + \tfrac{1}{2} \mathrm{tr}[(\Sigma_1^{-1}\Sigma_2 - I)^2]$$

ossia la distanza di Mahalanobis tra le medie più un termine che misura la divergenza tra le matrici di covarianza.

E.3 La congettura fenomenica

La congettura avanzata nella Parte Quinta del saggio può essere formalizzata nel modo seguente. Sia M lo spazio dei parametri del modello generativo del sistema cognitivo, dotato della metrica di Fisher-Rao g. Sia Q lo spazio fenomenico — lo spazio dei qualia — dotato di una distanza d_Q misurata mediante giudizi di somiglianza percettiva (scaling multidimensionale). Sia $\varphi : M \to Q$ la funzione che associa a ogni configurazione di parametri del modello generativo il quale corrispondente.

La congettura afferma che φ è un'isometria approssimata tra (M, g) e (Q, d_Q):

$$d_Q(\varphi(\theta_1), \varphi(\theta_2)) \approx \lambda \cdot d_{FR}(\theta_1, \theta_2)$$

per qualche costante di proporzionalità $\lambda > 0$ che dipende dalle unità di misura delle due distanze.

Le predizioni verificabili che ne derivano sono le seguenti. Nel dominio visivo dei colori, la distanza fenomenica tra due colori (misurata psicofisicamente tramite scaling multidimensionale) dovrebbe essere proporzionale alla distanza di Fisher-Rao tra le distribuzioni del modello generativo visivo corrispondenti ai due colori. Nel dominio uditivo, la distanza percettiva tra altezze o timbri dovrebbe corrispondere alla distanza di Fisher-Rao nello spazio dei parametri del modello generativo uditivo. Nel dominio somatosensoriale, le differenze qualitative tra tipi di dolore (acuto vs. cronico, bruciante vs. pressorio) dovrebbero riflettersi in distanze di Fisher-Rao corrispondenti nello spazio del modello interoceptivo.

La verifica di questa congettura richiede: (a) la stima dei parametri del modello generativo per stimoli fenomenicamente discriminati; (b) il calcolo della distanza di Fisher-Rao tra questi parametri; (c) la misura psicofisica delle distanze fenomeniche; (d) il test di correlazione tra le due misure. Questo programma di ricerca è tecnicamente impegnativo ma non in linea di principio impossibile.

Appendice F — Note sulla Dynamic Causal Modelling

F.1 Il problema della connettività effettiva

La Dynamic Causal Modelling (DCM) è una famiglia di metodi per l'inferenza sulla struttura causale di sistemi di equazioni differenziali a partire da dati di neuroimmagine. Il problema che affronta è il seguente: dati i segnali BOLD (Blood Oxygen Level Dependent) di N regioni cerebrali osservati durante un compito, inferire la struttura delle connessioni causali tra le regioni e come queste connessioni sono modulate dal contesto sperimentale.

F.2 Il modello neurale

Il modello neurale della DCM specifica l'evoluzione temporale dell'attività neurale z(t) come:

$$dz/dt = (A + \sum_j u_j B^j) z + C u$$

dove A è la matrice di connettività intrinseca (connessioni presenti anche in assenza di stimoli), B^j è la matrice di modulazione contestuale per il j-esimo fattore sperimentale, C è la matrice di guadagno dell'input, e u è il vettore degli input sperimentali. La risposta BOLD è generata da un modello neurovascolare (equazioni di Balloon-Windkessel) che traduce l'attività neurale in segnale BOLD.

F.3 L'inferenza bayesiana sui parametri

I parametri $\theta = (A, B, C)$ sono inferiti usando l'inferenza variazionale (solitamente l'approssimazione di Laplace) sulla verosimiglianza del segnale BOLD osservato dato il modello. L'energia libera variacionale F è il funzionale minimizzato:

$$F(q, y) = KL[q(\theta) \;//\; p(\theta)] - \mathbb{E}_q[\ln p(y \mid \theta)]$$

dove $p(\theta)$ è il priore sui parametri e $p(y \mid \theta)$ è la verosimiglianza dei dati BOLD. Il margine di evidenza del modello $-\ln p(y)$ è approssimato da $-F$ al minimo variazionale, e viene usato per il confronto bayesiano tra modelli con diverse strutture di connettività.

F.4 Connessione con il predictive coding

La DCM fornisce un metodo per verificare empiricamente le predizioni del predictive coding sulla struttura delle connessioni feedforward e feedback. In particolare, studi di Friston, Kiebel e collaboratori hanno usato la DCM per mostrare che le connessioni discendenti nella gerarchia visiva mostrano caratteristiche di modulazione compatibili con il trasporto di predizioni, mentre le connessioni ascendenti mostrano caratteristiche compatibili con il trasporto di errori di predizione — fornendo un supporto metodologicamente robusto, anche se non definitivo, all'architettura del predictive coding.

Appendice G — Glossario dei termini tecnici

Approssimazione di Laplace: Approssimazione di una distribuzione di probabilità arbitraria con una distribuzione gaussiana centrata sulla sua moda. Semplifica l'inferenza variazionale rendendola analiticamente trattabile.

Confine di Markov: Insieme di nodi in una rete bayesiana (o insieme di variabili in un sistema stocastico) la cui specificazione rende un nodo condizionalmente indipendente da tutti gli altri. Nei sistemi biologici, corrisponde all'interfaccia sensorimotrice tra organismo e ambiente.

Divergenza di Kullback-Leibler KL[q // p]: Misura di dissomiglianza tra due distribuzioni di probabilità q e p, definita come $\mathbb{E}_q[\ln(q/p)]$. Non è una distanza (non è simmetrica), ma è sempre ≥ 0.

Dynamic Causal Modelling (DCM): Metodo bayesiano per l'inferenza sulla struttura delle connessioni causali tra regioni cerebrali a partire da dati di neuroimmagine. Utilizza l'inferenza variazionale per stimare parametri di modelli di equazioni differenziali.

Energia libera attesa G(π): Funzionale su politiche che combina rischio (divergenza dagli stati preferiti) e ambiguità (incertezza epistemica residua). La sua minimizzazione guida la selezione delle azioni in modo da bilanciare motivazione strumentale e motivazione epistemica.

Energia libera variacionale F: Funzionale su distribuzioni di probabilità q e osservazioni o, definito come $\mathbb{E}_q[\ln q - \ln p(o, \psi)]$. È un limite superiore della sorpresa $-\ln p(o)$ e viene minimizzato nell'inferenza variazionale.

Errore di predizione: Discrepanza tra il valore predetto di una variabile (generato dal modello interno) e il suo valore effettivo (derivato dal segnale sensoriale o dal livello inferiore della gerarchia). Nel predictive coding, gli errori di predizione ponderati dalla precisione sono i segnali trasmessi verso l'alto nella gerarchia.

Inferenza attiva: Estensione del FEP in cui il sistema minimizza l'energia libera non soltanto aggiornando le credenze ma anche agendo sul mondo per rendere le osservazioni più compatibili con il modello generativo. Unifica percezione, azione e pianificazione in un unico formalismo.

Matrice di informazione di Fisher: Matrice che misura la quantità di informazione che una variabile casuale X porta riguardo al parametro θ di cui dipende la sua distribuzione. Definisce la struttura geometrica dello spazio delle distribuzioni (geometria dell'informazione di Amari).

Metrica di Fisher-Rao: Metrica riemanniana su uno spazio di distribuzioni di probabilità definita dalla matrice di informazione di Fisher. La distanza geodetica secondo questa metrica misura la distinguibilità statistica tra distribuzioni vicine.

Mismatch Negativity (MMN): Componente negativa del potenziale evento-correlato che emerge circa 100-200 ms dopo la presentazione di uno stimolo deviante in una sequenza di

stimoli standard. Interpretata come il correlato neurale dell'errore di predizione uditivo nel framework del predictive coding.

Modello generativo: Distribuzione di probabilità congiunta $p(o, \psi)$ su osservazioni o e stati nascosti ψ che codifica le assunzioni del sistema sulla struttura del mondo. La percezione è inferenza inversa su questo modello: dati o, inferire ψ.

Precisione: Inverso della varianza di una distribuzione gaussiana, o generalizzazione matriciale (matrice di precisione = inverso della matrice di covarianza). Nel FEP, la precisione pesa gli errori di predizione: alta precisione implica che gli errori vengano trattati come fortemente significativi.

Predictive coding: Architettura computazionale gerarchica in cui ogni livello genera predizioni sui livelli inferiori e riceve in cambio soltanto gli errori di predizione. Derivabile come implementazione dell'inferenza variazionale gaussiana sotto un modello generativo gerarchico.

Sorpresa (surprisal): $-\ln p(o)$. Misura quanto una osservazione o è inattesa sotto la distribuzione marginale $p(o)$ del modello generativo. La minimizzazione della sorpresa media è equivalente alla sopravvivenza in una nicchia ecologica.

Bibliografia

Opere fondative del FEP

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B*, 360(1456), 815–836.

Friston, K. (2008). Hierarchical models in the brain. *PLOS Computational Biology*, 4(11), e1000211.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, 13(7), 293–301.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, 10(86), 20130475.

Friston, K., Kilner, J., e Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1–3), 70–87.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., e Penny, W. (2007). Variational free energy and the Laplace approximation. *NeuroImage*, 34(1), 220–234.

Friston, K., e Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B*, 364(1521), 1211–1221.

Friston, K., Daunizeau, J., Kilner, J., e Kiebel, S. J. (2010). Action and behavior: a free-energy formulation. *Biological Cybernetics*, 102(3), 227–260.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., e Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, 6(4), 187–214.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., e Pezzulo, G. (2017). Active inference: a process theory. *Neural Computation*, 29(1), 1–49.

Predictive coding e neuroscienze computazionali

Rao, R. P. N., e Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79–87.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204.

Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., e Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4), 695–711.

Millidge, B., Tschantz, A., e Buckley, C. L. (2020). Predictive coding approximates backprop along arbitrary computation graphs. *arXiv preprint* arXiv:2006.04182.

Auksztulewicz, R., e Friston, K. (2016). Repetition suppression and its contextual determinants in predictive coding. *Cortex*, 80, 125–140.

Inferenza attiva e psicopatologia computazionale

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., e Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, 4, 47.

Corlett, P. R., Honey, G. D., Krystal, J. H., e Fletcher, P. C. (2011). Glutamatergic model psychoses: prediction error, learning, and inference. *Neuropsychopharmacology*, 36(1), 294–315.

Fletcher, P. C., e Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10(1), 48–58.

Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A. E., Paliwal, S., Gard, T., Tittgemeyer, M., Fleming, S. M., Haker, H., McAlonan, G. M., e Petzschner, F. H. (2016). Allostatic self-efficacy: a metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in Human Neuroscience*, 10, 550.

Lawson, R. P., Rees, G., e Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in Human Neuroscience*, 8, 302.

Kapur, S. (2003). Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. *American Journal of Psychiatry*, 160(1), 13–23.

Filosofia della mente e coscienza

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450.

Levine, J. (1983). Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly*, 64(4), 354–361.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227–247.

Jackson, F. (1982). Epiphenomenal qualia. *The Philosophical Quarterly*, 32(127), 127–136.

Kripke, S. A. (1980). *Naming and Necessity*. Harvard University Press.

Goff, P. (2017). *Consciousness and Fundamental Reality*. Oxford University Press.

Goff, P. (2019). *Galileo's Error: Foundations for a New Science of Mind*. Pantheon Books.

Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, 23(11–12), 11–39.

Strawson, G. (2006). Realistic monism: why physicalism entails panpsychism. *Journal of Consciousness Studies*, 13(10–11), 3–31.

Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. MIT Press.

Enattivismo e cognizione incarnata

Varela, F. J., Thompson, E., e Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Harvard University Press.

Noë, A. (2004). *Action in Perception*. MIT Press.

Teoria computazionale della mente, modularità, connessionismo

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Fodor, J. A. (1983). *The Modularity of Mind*. MIT Press.

Fodor, J. A., e Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition*, 28(1–2), 3–71.

Rumelhart, D. E., e McClelland, J. L. (a cura di) (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (2 voll.). MIT Press.

Geometria dell'informazione e statistica

Amari, S. (1985). *Differential-Geometrical Methods in Statistics*. Springer.

Amari, S., e Nagaoka, H. (2000). *Methods of Information Geometry*. American Mathematical Society.

Cover, T. M., e Thomas, J. A. (2006). *Elements of Information Theory* (2ª ed.). Wiley.

Testi storici e filosofici di riferimento

Helmholtz, H. von (1867). *Handbuch der physiologischen Optik*. Leopold Voss.

Ryle, G. (1949). *The Concept of Mind*. Hutchinson.

Davidson, D. (1980). *Essays on Actions and Events*. Oxford University Press.

Kim, J. (1998). *Mind in a Physical World*. MIT Press.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman.

Russell, B. (1927). *The Analysis of Matter*. Kegan Paul.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Merleau-Ponty, M. (1945). *Phénoménologie de la perception*. Gallimard.